UNIVERSITY OF TORINO

Ph.D. in Modeling and Data Science

Final dissertation

Exploring Algorithms for Sentiment Analysis in Medical Web Pages, and Sentiment-Preserving Extractive Summarization

Supervisor: Prof. Luca Anselma
Co-supervisor: Prof. Alessandro Mazzei

Candidate: Md. Murad Hossain

ACADEMIC YEAR 2023/2024

Summary

This thesis conducts a thorough application of natural language processing, with a specific focus on sentiment analysis in medical web content. Additionally, it explores techniques for creating extractive summaries that effectively capture and preserve sentiments.

Firstly, our research delves into the emotional landscape of medical web content through sentiment analysis by an exploratory Study for the development of the SentiTextRank algorithm. Through careful examination of emotional patterns, we particularly focus on the role of disgust in discerning between different musculoskeletal conditions. During our investigation, we examine psychological differences and potential impacts on patient experiences. This insight improves our understanding of how emotional factors affect the dissemination of health-related information.

In this study, we have tried to develop two algorithms for extractive summarization: SentiTextRank Version 1 and Version 2. Version 1 prioritized content similarity through techniques like TextRank to rank sentences based on their similarity scores. Version 2 improved upon this by integrating emotion similarity, utilizing an emotion analysis module to extract emotional features and combine them with content similarity. This integration allowed for summaries that preserved both the factual content and emotional nuances of the original text, marking a significant advancement in the SentiTextRank approach.

Additionally, we endeavor to refine sentiment-preserving extractive summarization techniques, with a particular emphasis on SentiTextRank's evolution and the delicate balance between emotional resonance and factual accuracy in condensed summaries. This thesis not only uncovers strategies for navigating emotional intricacies in health-related information but also advances methods for crafting concise yet emotionally resonant summaries, aiming to enhance communication interfaces and information representations in natural language processing. While acknowledging ongoing

challenges, our work underscores the potential for synergistic collaboration among these research areas, highlighting their shared potential for business applications and their capacity to collectively enhance customer care in a more comprehensive manner.

Through comprehensive inquiries, our thesis presents innovative findings. It exposes strategies to uncover emotional intricacies influencing health-related information and enhances methods for creating concise yet emotionally resonant summaries. The research aims to enhance communication interfaces and information representations in natural language processing, focusing on adaptability, emotional impact, and contextual relevance. Despite some less favorable results, ongoing efforts continue to explore these aspects while recognizing challenges.

Acknowledgements

I owe a debt of gratitude to my co-supervisor, professor Alessandro Mazzei, and supervisor, professor Luca Anselma, for their unwavering support, priceless counsel, and profound knowledge, all of which have greatly influenced the direction of this study. Their academic wisdom, support, and guidance have been invaluable in helping me navigate the challenges of this PhD project.

I extend my sincere appreciation to our program Coordinator Prof.Laura Sacerdote for her unwavering encouragement, insightful feedback, and dedication to fostering academic excellence. Her support significantly contributed to the refinement and success of my research.

My sincere gratitude also extends to the University of Turin's faculty and personnel, whose dedication to academic quality created a stimulating learning atmosphere. Their continuous support has been invaluable throughout this academic pursuit.

To my colleagues - thank you for your camaraderie; our stimulating discussions enriched my academic journey!

Lastly, I would like to express my gratitude to my family, especially to my mother, wife, and daughter, for their extraordinary sacrifices and consistent encouragement. Their unwavering support has been a vital source of inspiration for me the entire way.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and motivation

This doctoral dissertation embarks on an ambitious exploration of two key areas in the rapidly emerging fields of natural language processing and computational linguistics that hold great promise for improving our understanding and application of sentiment analysis in medical web pages and summarization. The disciplines of sentiment analysis in medical content through an exploratory study for the development of SentiTextRank, and extractive summarization preserving sentiment, all significant aspects of linguistics, are the subjects of our investigation.

The Internet has transformed the way people access health-related information, making it a rich source of knowledge. However, the emotional dimension of medical content, often understudied, is gaining prominence due to its potential influence on user perceptions and decision making. With an emphasis on chronic health issues like back pain, this dissertation aims to expand on our knowledge of sentiment analysis in the context of medical web content. This is driven by two factors. First, the widespread occurrence of long-term health issues and the importance of emotions in determining health results require a thorough investigation into sentiment analysis. Second, the abundance of online health information emphasizes how important it is to recognize emotional patterns and how they affect user behavior. Our goal in examining this emotional aspect is to offer valuable insights for providers and consumers of health information.

This dissertation also introduces sentiment-preserving extractive summarization, emphasizing the preservation of emotional nuances alongside core information in text summarization. To enhance the emotional resonance and the quality of the summarization, we have created "SentiTextRank"

an emotional variant of the TextRank algorithm. Many datasets have been used with it, including news, DUC2001, DUC2004, wikihow, podcast transcripts, tweets, blog posts, Bbc articles, CNN news, and Hippocorpus stories. SentiTextRank is compared to the original TextRank using cosine similarity or content overlap similarity and BERT sentence embeddings. SentiTextRank categorizes sentences by emotional content using SenticNet's vocabulary.

Our target was to generate extractive summaries using our proposed SentiTextRank algorithm for version 1 and Version 2. But in real situations, in most of the cases, we got abstractive format reference summaries for all datasets. For evaluating our generated summary through different metrics in perspective of reference summary, we need extractive reference summary also. That is why We created extractive reference summaries using maximizing Rouge scores. Our evaluation metrics include emotional distance concerning gold summary and original text, Rouge-L-F1, Bert-F1 score, Cosine similarity, Mover score, and Pyramid score to assess how well the generated summary is. Evaluations involve contrasting our summaries with reference summaries and assessing their effectiveness in relation to about other approaches.

In addition to advancing text analysis and sentiment-preserving summarization techniques, this doctoral dissertation aims to strengthen the interrelation between the explored areas, recognizing their shared potential for practical applications, particularly in business contexts. While the focus is on sentiment analysis in medical web content and the development of sentiment-preserving summarization algorithms, the implications of these advancements extend beyond their individual domains. By acknowledging the interconnectedness of these research areas, we aim to highlight their collective impact on improving customer care and enhancing user experiences. Through a comprehensive exploration of emotional aspects in medical content and innovative summarization techniques, this dissertation seeks to bridge the gap between theory and practice, offering valuable insights for various application domains.

In summary, this doctoral dissertation represents a multifaceted project aimed at advancing text analysis through sentiment analysis in medical web content and sentiment-preserving extractive summarization across diverse domains.

1.2 Problem statement

Among the multifaceted fields of computational linguistics and natural language processing, this extensive research project encompasses two related domains: sentiment analysis in medical web pages, and extractive summarization preserving sentiment. Each of these research areas poses distinct, yet interrelated, challenges, demanding exploration and innovation.

The digital sphere has emerged as a crucial avenue for obtaining health-related information. Our emphasis is on comprehending the emotional content present in medical web pages, particularly those on chronic conditions such as back pain. The challenge at hand encompasses two aspects: the emotional dimension of medical web content remains under-explored and there is insufficient understanding of the potential impact of emotions on user decision-making. Consequently, this study seeks to scrutinize the emotional content within medical web pages and explore its effect on user behavior within the domain of chronic conditions.

Text summarization is crucial in condensing large amounts of text into concise summaries. However, current techniques tend to overlook the emotional context of the text, resulting in summaries that lack emotional nuances and resonance. This issue becomes more pronounced when applied to diverse data sets from fields such as news, medical content, blogs, stories, and scientific publications. We present a unique method for emotional text summary in this study that produces emotionally resonant summaries while taking into account the text's emotional aspects.

This doctoral thesis aims to tackle these complex challenges by investigating new algorithms and methodologies in each of these areas. To address the lack of emotional content preservation in summaries, we present the novel SentiTextRank method for text summarization. The main challenge is to create holistic solutions that, identify emotional patterns in medical web content, and produce emotionally resonant summaries but also acknowledge the interconnected nature of these issues and handle diverse datasets used during this process. The ultimate objective is to contribute to the advancement of natural language processing and computational linguistics by enhancing user experiences and offering valuable insights for various application domains. Overall, this thesis seeks to investigate and design all-encompassing solutions to the problems of recognizing emotional patterns in medical web content and generating emotionally rich summaries.

1.3 Research objectives

This doctoral thesis encompasses a multi-faceted exploration of two critical areas: Sentiment Analysis in Medical Web Pages, and Sentiment-Preserving Extractive Summarization. This extensive study aims to investigate the following research objectives:

1.3.1 Research Objectives for sentiment analysis in medical web pages

1. Gather a diverse and substantial data set from multiple languages, with a specific emphasis on English, ensuring data consistency for sentiment analysis.

2. Conduct sentiment analysis on web pages related to specific medical conditions, with a primary focus on chronic back pain, to uncover the underlying emotional patterns within the content. Also, represent these emotional patterns using histograms to create distinctive "emotional fingerprints" for each health condition.

3. Conduct a comparison of emotional patterns across different health conditions to assess significant differences in expressed emotions, both graphically and statistically.

4. Employ machine learning algorithms, specifically linear Support Vector Machines (SVM), to classify Internet documents based on their emotional content, striving to achieve classification accuracy exceeding 90%.

5. Thoroughly analyze the results of machine learning classification, addressing implications and limitations, especially the influence of emotional content on user perceptions and decision-making.

1.3.2 Research Objectives for Sentiment-Preserving Extractive Summarization

1. Evaluate various summarization techniques across different domains, assessing their ability to capture both essential information and emotional nuances within the text.

2. Develop an emotional variant summarization technique, SentiTextRank, which introduces sentiment analysis into the summarization process.

The goal of this innovation is to produce summaries that accurately capture the text's emotional tone.

3. Experiment with different weight assignments for factors such as cosine similarity, content overlap similarity, and emotional scores to observe their impact on the quality of the produced summaries and the preservation of emotional sentiments.

4. Develop algorithms that prioritize sentences for summary generation based on compression ratios, ensuring the selection of emotionally rich sentences from the text.

5. Rigorously assess the effectiveness of SentiTextRank and other summarization methods by using standard metrics, user preferences, and domain-specific evaluations, the generated summaries are compared with reference summaries.

Overall, the purpose of this thesis is to advance the field of natural language processing and computational linguistics by enhancing summarization techniques with emotional nuances. The research also seeks to investigate user intentions, sentiment analysis in medical web content, and sentiment-preserving extractive summarization.

1.4 Scope and limitations

In this research project, we examine two related areas:

Analyzing Sentiments in Medical Web Pages, and Preserving Sentiments in Extractive Summarization. The study covers a broad array of domains and datasets. First, we analyze sentiments in medical web pages to uncover emotional patterns across different health conditions and tried to assess significant differences in expressed emotions through the exploratory study of the development of SentiTextRank.

Finally, we explore the challenge of preserving sentiment in extractive summarization by experimenting with various weight assignments for factors like cosine similarity, content overlap similarity, and emotional scores.

We broaden our focus to health-related information available on the internet, notably in the area of Medical Web Pages Sentiment Analysis. Our research concentrates on web pages related to medical conditions, particularly chronic back pain. The aim is a cross-linguistic analysis that seeks to reveal emotional patterns in web content and evaluate their potential

influence on user perceptions and decision-making. Additionally, we strive to enhance the effectiveness of extractive summarization techniques by retaining sentiments in the generated summaries.

Our research is centered on enhancing text summarization while retaining the emotional content present in the original text. We aim to analyze various types of text, including news articles, medical documents, stories, and blog posts. Our plan involves comparing different summarization methods such as original TextRank, an emotional variant of TextRank(SentiTextRank version 1 and Version 2) with cosine similarity, and Bert sentence embedding. Furthermore, for carrying out the SentiTextRank experiment in which we use SenticNet to create summaries while taking emotion categories into account. Utilizing metrics, we will assess the level of quality of our summaries including Rouge-L, Bert-score, cosine similarity/Content overlap similarity metric, Mover score, pyramid score, and emotional distance metrics. In addition to this analysis, to aid in decision-making about the inclusion of sentences into the summary, we will examine compression ratios.

Nevertheless, it is imperative to recognize the constraints of our study. The effectiveness of the algorithms and models developed in this research may vary depending on the specific domain and data set being analyzed. The features and type of the data being analyzed determine how well these strategies work.

Sentiment analysis methods like SenticNet's performance are crucial to the effectiveness of sentiment-aware strategies and the level of accuracy of the summaries they produce. Inaccuracies in sentiment analysis can impact the emotional resonance of the summarization process.

Our research primarily focuses on enhancing extractive summarization techniques with sentiment preservation. Our specific choice of using extractive summarization was influenced by our utilization of a modified TextRank algorithm (SentiTextRank), which inherently generates extractive summaries. Additionally, the available reference summaries in the dataset were not consistently in an extractive format. To address this, we generated extractive reference summaries by maximizing the Rouge score based on the given abstractive reference summaries. These factors guided our decision to use extractive summarization methods rather than abstractive method.

The determination of the most suitable summarization method can be subjective and depends on individual preferences. Different stakeholders may have varying perceptions of the quality and effectiveness of summarization techniques.

In evaluating the generated summaries, one difficulty we face is the lack of extractive reference summaries for all methods. To address this limitation, we take a practical approach by creating extractive reference summaries using estimated Rouge scores from given abstractive reference summaries. While this method helps with evaluation, it could introduce potential differences in summary quality and accuracy.

Notwithstanding these drawbacks, our main goal is to advance the domains of computational linguistics and natural language processing, and user-centric information processing by integrating sentiment and emotions. This integration results in more accurate and emotionally meaningful summaries across diverse domains. The goal of this research is to assess and improve the quality of generated summaries using different evaluation metrics that provide a robust framework for evaluating both quality and emotional resonance. In conclusion, our research aims to enhance text summarization techniques by retaining sentiments in the generated summaries.

1.5 Thesis outline

This thesis comprises four chapters, each dedicated to a distinct research area.

Chapter 1:"Introduction", serves as the foundational chapter that establishes the research's context of background and motivation, articulates the problem statement, and delineates the objectives of the study. It prepares the reader for the subsequent research areas.

Chapter 2: "Sentiment Analysis in Medical Web Pages", expands the investigation to consider the overlap between health information and emotions in online content. It addresses the influence of the Internet on accessing health-related information, introduces the research hypothesis, details data collection methods, and examines emotional content analysis on web pages. Furthermore, this chapter discusses implementing machine learning classification, shares results and implications and outlines future avenues for this field.

Chapter 3: "Sentiment-Preserving Extractive Summarization", covers a detailed exploration of the field, starting with definitions and the critical significance in the domain. The chapter discusses various methodologies and algorithms essential to extractive summarization, including the TextRank and PageRank algorithms. It also explores advanced techniques like

BERT Sentence Embedding and Cosine similarity/Content Overlap similarity within the summarization landscape. In this section, SentiTextRank is introduced, and its various versions are explored to highlight their strengths and limitations. It also places these methods in the context of other summarization tools before delving into dataset selection and the experimental setup. The chapter then provides an overview of evaluation metrics for summarization tasks, followed by a detailed analysis of the experimental results for both versions of SentiTextRank. These results are thoroughly examined, leading to valuable insights discussed in subsequent sections and concluding with future directions for sentiment-preserving extractive summarization exploration.

Chapter 4: The main Conclusion , their importance, and their applications are outlined in the "Conclusions" section. This chapter identifies the study's shortcomings and suggests directions for further investigation.

The thesis concludes with three appendices: In Appendix A, Side PhD work: Anticipating User Intentions in Dialogue Systems, In Appendix B named as thesis Appendices where we include our published articles and in Appendix C we have included our repository link for related code of thesis.

Chapter 2

Sentiment Analysis in Medical Web Pages

2.1 Introduction

Sentiment analysis, an essential component of natural language processing, is reshaping our understanding of how online information influences emotions and decision-making. It is a useful resource for understanding the viewpoints and emotions stated on websites pertaining to medicine.

The intricate relationship between chronic pain, such as back pain, and emotions calls for a deeper understanding of the affective information present in medical web pages. This study focuses on the sentiment analysis of online content related to spine and great joint pathologies, exploring the emotional underpinnings that may influence the patient's perception and behaviors towards their condition.

With the increasing reliance on the internet for health-related information, the significance of digital emotional fingerprints cannot be overstated; they potentially integrate into the biopsychosocial framework of patients' experiences.

Machine learning algorithms employed in this work aim to discern patterns within these fingerprints, potentially enabling the prediction of web page topics based on their emotional content, thereby shedding light on the integration of a patient's biopsychosocial ecosystem into digital affective "successful content", with implications for understanding the emergence of chronic pain and endorsing health-relevant behaviors.

This chapter delves into the background and significance of sentiment analysis in medical web pages, explaining related literature with combining a review of methods and tools with a focus on machine learning's impact

on healthcare.

2.1.1 Background and significance of sentiment analysis in the medical domain

Due to its ability to combine the complex distinctions of human emotions with the computational power of machine learning, sentiment analysis has gained significant importance in the healthcare sector. In healthcare, where patient narratives are abundant, sentiment analysis serves as a crucial tool to uncover patient-centered insights that conventional numerical data may overlook.

Sentiment analysis plays a crucial role in medicine, providing a data-driven method for understanding the complex emotions that surround health. With the increasing availability of online medical information, integrating sentiment analysis can improve patient engagement, shape health communication strategies, and enrich our comprehension of patient experiences. The incorporation of sentiment analysis into health informatics has the potential to transform healthcare communication by prioritizing a patient-centered approach in our constantly evolving digital healthcare landscape.

Additionally, sentiment analysis assesses the emotional content of medical web pages, which is especially significant given that the Internet is a major source of health information. The sentiments expressed in digital health content influence patient perspectives and behaviors, impacting health outcomes and the dissemination of medical knowledge.

Ultimately, more patient-centric and emotionally intelligent healthcare practices result from the integration of remarkable improvements in machine learning and sentiment analysis, which function as an inspiration and redefine how we comprehend emotions in healthcare settings.

Numerous research studies demonstrate the importance of sentiment analysis, ranging from assisting in the early identification of public health risks by monitoring social responses to assessing the emotional tone of medical websites. This is particularly crucial considering the Internet's significance as a key provider of health-related information. To further clarify, sentiment analysis is positioned to influence the future of healthcare communication, providing distinct perspectives into patient attitudes and actions with profound implications for health outcomes and the sharing of medical information. The section that follows will take a close look at the articles that contribute to the understanding of sentiment analysis's significance in the healthcare sector.

The authors of [190] delves into sentiment analysis methodologies and tools utilized for extracting sentiments from textual data, particularly focusing on social networks as valuable sources. The researchers classify current tools by factors like technology and interoperability, test them in practical situations, and then discuss their limitations and the necessity for improvements.

Building upon this foundation, We explore research on medical forums, drug reviews, and clinical narratives, examining sentiments related to healthcare experiences.

The work from authors [89] use supervised learning and lexicon-based techniques to examine opinions expressed in medical forums, namely on doctors and pharmaceuticals. This study sheds light on the complexity of language within drug reviews.

The subsequent investigation in the authors of[34] conducts a comprehensive study on sentiment analysis in patient-authored content in online health forums. Positive, negative, and neutral sentiments are applied to health information, and different machine-learning algorithms and feature sets are tested. The findings emphasize the importance of factual statements with inherent sentiment polarity ("polar facts") in online health forums.

Our exploration extends to sentiment analysis in drug reviews in the work from authors [68] addressing challenges such as the scarcity of annotated data and variability in user language. Logistic regression models using lexical features predict sentiments related to patient satisfaction, side effects, and drug effectiveness. This study shows how domain-specific language affects model performance and proposes that cross-data sentiment analysis could be enhanced by using a bigger training set.

The authors of [48] discussed the application of sentiment analysis in clinical documents, highlighting differences between clinical narratives and medical social media. Their quantitative analysis, based on a domain-specific corpus, emphasizes understanding both explicit and implicit sentiments in clinical texts for patient health status and treatment outcomes.

Further enriching our understanding of sentiment analysis in healthcare, a study on online reviews for spine surgeons in the work from authors [163] underscores the significance of pain management and positive interpersonal behaviors in influencing patient satisfaction. The authors stress the ongoing need for research to assist surgeons in improving their practices and online presence, emphasizing the crucial role of sentiment analysis in comprehending patient experiences and feedback.

The subsequent investigation introduces a linguistic approach to sentiment analysis in user-generated drug reviews the authors of [123], addressing limitations associated with traditional machine learning methods. This work outperforms SVM techniques in precision, recall, F-score, and accuracy by segmenting sentences into clauses and using specialized lexicons to assign sentiment scores.

The work focused on text mining within medical forums places a lot of emphasis on sentiment analysis of talks linked to hearing loss in the work from authors [12]. The authors demonstrate how classifiers like Naïve Bayes may greatly enhance performance by reliably classifying forum posts using Machine Learning approaches such as SVM, Naïve Bayes, and Logistic Regression.

Social media platforms are included in the investigation, especially as they relate to public health surveillance in the authors of[11].The Sentiment Analysis as a Service (SAaaS) framework, designed for sentiment extraction and analysis, is presented as a dynamic service composition mechanism to address data characteristics and noise in social media information. The authors highlight the value of sentiment analysis in public health by demonstrating how well the SAaaS framework monitors flu epidemics.

A thorough assessment of earlier research on sentiment analysis is encountered by the work from authors [143], which highlights a research gap regarding the application of sentiment analysis to categorize patients' emotional states in healthcare. This observation leads to the introduction of a context-specific tool, SentiHealth-Cancer, aimed at improving mood detection in cancer patients through Portuguese language posts from Brazilian patient communities.

Using sentiment analysis to mine health social media in the authors of [181], the authors examine user-generated content in online health individuals. Their proposed framework, comprising stages such as extracting medical terms, enhancing the Latent Dirichlet Allocation algorithm, and conducting sentiment analysis, offers valuable findings for patients, caregivers, and doctors in the healthcare domain.

The researchers discussed levels of sentiment analysis, best practices for identification and classification, data collection methods from online sources including social media and forums, importance of feature extraction. They also reviewed sentiment analysis tasks across various domains such as customer feedback in hotel reviews or stock market analysis. The survey addressed challenges faced in sentiment analysis like handling sarcasm and

regional linguistic variations while providing a comparative analysis of different approaches to sentiment analysis. In order to properly evaluate user sentiments in a variety of circumstances, it concluded by outlining the current status of research in this area and emphasizing the need for additional developments in sentiment analysis approaches in the work from authors [175].

The authors of [133] conducted a review of 12 studies on sentiment analysis in assessing public opinion on health technologies shared through various social media platforms. Despite challenges, sentiment analysis was recognized for its ability to rapidly collect large data sets reflecting public sentiment and informing further analysis. The findings suggest significant potential for utilizing social media data in health technology assessments and decision-making processes while incorporating patient perspectives without conflicts of interest.

The work from authors [106] examines how individuals with psychological disorders and those with physiological diseases differ in their discussion themes and emotional reactions in online health groups. It highlights disparities in conversation themes and a higher prevalence of intense emotions among individuals facing psychological disorders. The research suggests potential implications for improving medical services, creating specialized internet-based health forums, tailoring interventions to address specific patient needs, considering constraints like the informal nature of platforms such as Baidu Post Bar, and identifying future areas for investigation.

The authors of [119] developed a medical lexicon named WordNet of Medical Events to automate the extraction and classification of medical concepts, including categorization and sentiment analysis. They validated their systems using supervised machine learning classifiers and showcased applications in recommendation systems and medical question answering.

The work from authors [71] proposes a novel approach to biomedical sentiment analysis, integrating deep associative learning with neural networks. Their model leverages unsupervised learning and associative memory to effectively capture the nuances of medical concepts and sentiments in large-scale patient narrative data, outperforming existing baselines in detecting medical concepts and their associated sentiments.

A summary of sentiment analysis techniques in the healthcare industry is given in the excerpt, which also highlights the significance of specialized methodologies given the influence that medical conditions have on patients'

quality of life. It discusses techniques, domain dependence, and tools employed in this context to improve healthcare services and patient experience through insights from social media in the authors of [1].

By examining the relationship between official health data and social media data, the article investigates the application of machine learning and sentiment analysis in the prediction of disease outbreaks. It examines different ML algorithms for healthcare prediction models, identifies research gaps, and emphasizes technology's critical role in addressing epidemic challenges in the work from authors [155].

The authors of [162] studied sentiments about telemedicine on Twitter, comparing different sentiment analysis methods. They found that lexical and semantic-based methods were more accurate when training datasets were not evenly distributed. Domain-specific language influenced the performance of prediction methods in demographic distribution related to health topics showcased by Twitter analysis along with its limitations.

A scoping review in the work from authors [153] examined 10 studies from 2002 to 2019, revealing common themes like the predominant use of Twitter as the analyzed platform, sentiment analysis for assessing public perception towards medications and adverse drug reactions, and how news media sources affect the sentiment on social media. The study highlighted difficulties unique to sentiment analysis in the healthcare industry and recommended more research employing machine learning and lexicon-based approaches.

The authors of [34] looked at sentiment analysis in forums related to e-health, including classifying patient-authored content and evaluating textual representation techniques. The researchers created the eDiseases dataset, analyzed polar facts prevalence, and implemented machine learning algorithms to predict content polarity. They concluded by emphasizing accurate sentiment analysis' potential to enhance online health information accessibility and outlined future work on identifying drugs and treatments as well as detecting adverse drug events in social media posts.

Sentiment analysis in the healthcare industry provides valuable patient viewpoints and makes a substantial contribution to understanding public opinion around health technologies on social media. Ongoing advancements aim to increase the accuracy of comprehending complex medical concepts and patient emotions, one such improvement being the application of advanced associative learning techniques. This has enormous potential to improve patient outcomes and healthcare services. Sentiment analysis is

a valuable tool in healthcare communication because it may be used to interpret patient attitudes expressed online, influence public opinion, and strengthen relationships between physicians and patients. The constant search of improving these techniques shows a dedication to fully use sentiment analysis's potential in a range of healthcare contexts.

2.1.2 Related work on sentiment analysis and classification

In the expansive field of sentiment analysis and classification, researchers are exploring detailed insights from textual data found in social media posts, product reviews, and more. This journey has presented numerous challenges and opportunities, leading scholars to seek ways to improve the accuracy, depth, and applicability of sentiment analysis techniques. From addressing fundamental research gaps and challenges to harnessing advancements in techniques and algorithms, the landscape of sentiment analysis research is diverse. Sentiment analysis has applications in various industries and contexts, such as healthcare discussions, social media trends, and consumer preferences. Evaluating different sentiment analysis methods has been crucial for understanding their effectiveness and applicability.

Moreover, the combination of semantic analysis and feature extraction methods has made the way for advancements in sentiment analysis, empowering researchers to explore more deeply into the fundamental frameworks of textual data.

Additionally, domain-specific sentiment analysis has customized sentiment analysis tools to suit the distinct traits of various industries, leading to more accurate and contextually appropriate analyses.

The following section delves into a wide array of articles that address different facets of sentiment analysis and classification, encompassing gaps in research, methodological advancements, applications specific to certain domains, and more. Through these articles, the researchers provide valuable perspectives on the dynamic field of sentiment analysis research and its significant influence across diverse domains and disciplines.

There are many obstacles and difficulties in sentiment analysis that make it difficult to infer people's emotions from textual data. These complications might be thought of as barriers to comprehension, much like those that scientists must conquer to advance sentiment analysis. Luckily, scholars have written about these challenges and provided strategies for enhancing accuracy in sentiment analysis with some articles. Reviewing this literature will give valuable insights into the difficulties encountered by researchers

and their suggested remedies.

In natural language processing, sentiment analysis and classification have gained popularity, especially when it comes to examining emotions based on high-level emotion categories such as positive, neutral, and negative. A research gap that exists in sentiment analysis is the need for more advanced techniques to accurately classify sentiments beyond the basic emotion categories in the authors of [177].

Analyzing emotional states like happiness, anger, disgust, sadness, fear, and surprise, for example, is becoming more and more popular. These emotions offer a more nuanced understanding of sentiments and can provide valuable insights into people's opinions and behaviors. Another research gap in sentiment analysis is the limited scope of results provided by existing approaches. Most existing approaches in sentiment analysis offer basic sentiment classification in the positive-negative spectrum in the work from authors [97].

Advanced sentiment analysis methods, also known as beyond-polarity sentiment analysis, have not yet gained widespread use or full comprehension. These methods investigate intricate feelings such as joy, sadness, and disappointment, requiring advanced algorithms for accurate comprehension and categorization of emotions. The intricacies of human emotions continue to present a hurdle in the creation of universally effective advanced beyond-polarity sentiment analysis techniques by the authors of [131].

In sentiment analysis, researchers are constantly working to improve methods for understanding emotions in written text. This involves finding new ways to handle the complexities of textual data and enhancing the precision and comprehensiveness of sentiment analysis techniques. Studying some recent literature on these innovative approaches can provide valuable insights into their potential impact on sentiment analysis methods.

Sentiment analysis has seen an increase in the use of deep learning and natural language processing methods in recent years. With the use of neural networks, these methods can extract intricate patterns and characteristics from textual input, improving the accuracy of sentiment classification. Researchers have also started applying sentiment analysis to specific domains such as movie reviews, where the opinions of the users can provide valuable feedback on the quality of a film and its reaction. However, despite the progress made in sentiment analysis, there is still room for improvement in the work from authors [2].

Researchers are now exploring ways to improve sentiment analysis by

considering a wider range of emotions and sentiments beyond just positive, negative, and neutral. By incorporating more nuanced emotions, such as joy, surprise, and fear. The study of sentiment can offer a more thorough comprehension of people's opinions and behavior by the authors of [100].

In addition to improving the classification of emotions, another area of focus in sentiment analysis research is the scope of results provided by existing approaches. While basic sentiment analysis techniques offer a binary classification of positive or negative, more advanced techniques aim to provide results in terms of emotions, allowing for a deeper analysis of sentiments. Beyond basic positive or negative categorization, this beyond-polarity sentiment analysis can provide important insights into people's emotional states and reactions to different inputs and emotions. It can also provide the sentiment stated in a text broader context in the work from authors[69].

Applying sentiment analysis across several areas is similar to customizing a tool for a particular task. Every industry, be it social media, film reviews, healthcare, or education, offers different sentiment analysis opportunities and obstacles. We can learn a lot about sentiment analysis's real-world uses and the specialized methods required to extract insightful data by looking at how it is applied in these particular fields. Here, we explored some articles that showcase the various applications of sentiment analysis in specific domains and the lessons we can learn from them.

Sentiment analysis on social media platforms, such as Facebook, Instagram, and Twitter, is becoming more and more crucial to examining vast volumes of user-generated content. This can provide perceptions and opinions about the education of students.To classify a greater variety of emotions and increase the accuracy and comprehensiveness of sentiment analysis, researchers have proposed deep learning and natural language processing techniques by the authors of [17].

Furthermore, sentiment analysis also holds significant potential in the evaluation of educational webpages. In today's digital age, where online learning platforms and educational websites are increasingly prevalent, it is crucial to assess the effectiveness and user satisfaction of these platforms in the work from authors [91].

Through an analysis of user sentiment seen in comments, reviews, and feedback, administrators and educators can learn a great deal about the advantages and disadvantages of their webpages. By using this data, the websites' layout and content delivery may be enhanced, giving doctors a more customized and interesting learning experience by the authors of [36].

Machine learning and sentiment analysis techniques to investigate the relationship between a particular movie's ticket sales and its online reviews are used by researchers. Using online reviews, the researcher proposes that a simplified sentiment-aware autoregressive model may reliably predict box office sales. Fuzzy clustering, SVM Classifier, and tf-idf values are employed in document-level sentiment analysis to distinguish between positive and negative attitudes. They haven't, however, made much progress in identifying double negative expressions in the work from authors [124].

When studying sentiment analysis techniques, we explore the tools available for understanding textual emotions. Researchers have developed machine learning algorithms and deep learning architectures to handle sentiment analysis tasks effectively through the following articles. Understanding these techniques is important for grasping how sentiment analysis works and changes over time. Articles explain the different techniques used in sentiment analysis and their real-world applications.

Many sentiment analysis techniques have been developed to fill the research gap. Text documents are frequently classified using machine learning methods such as Naive Bayes classifiers and support vector machines, which identify positive and negative contents. Large amounts of text data can be efficiently analyzed by these algorithms to produce sentiment classifications by the authors of [102].

Using SentiWordNet, a lexical resource from the WordNet database, a semantic feature selection technique has been presented for opinion mining. In order to choose specific predicted terms, the method reduces the set of characteristics. To evaluate the feature selection techniques, experiments using Nave Bayes, FLR, and AdaBoost classifiers were carried out, and the results were compared in the work from authors [95].

Agarwal and Mittal investigate machine learning methods for extracting features, such as a semantic clustering approach and new bi-tagged features. They evaluate methods such as information gain, emphasize the importance of feature selection, and propose the mRMR approach as a way to reduce redundancy. They use preexisting data sets from a variety of domains to assess the algorithmic performance of support vector machines and boolean multinomial naive Bayes through ten-fold cross-validation by the authors of [5].

Techniques to increase the level of accuracy of sentiment analysis based on clustering are presented in a novel manner. For fine-grained sentiment analysis, the approach makes use of opposing opinions and non-opinion

processing, a redesigned voting mechanism, and an alternative distance measurement technique. The outcomes demonstrate the effectiveness of the clustering-based method, particularly in identifying neutral viewpoints. However, its drawbacks—such as the constraints of the k-means method and the size, shape, and balance of the data—have not been taken into account in the work from authors [99].

When researchers assess and contrast sentiment analysis techniques, it is necessary to take the role of evaluators in analyzing the merits and shortcomings of various approaches. By conducting thorough assessments and comparisons, their goal is to identify the most effective methods for different conditions and datasets. Examining literature that explores the assessment and comparison of sentiment analysis methods reveals detailed insights into this essential facet of research in sentiment analysis.

An automatic sentiment analysis method based on unsupervised ensemble learning was described by the authors. The approach consists of two stages: unsupervised learning with an ensemble of clustering classifiers that use a majority voting mechanism and contextual analysis with five processes. A modified k-means method serves as the basis classifier, and additional sentiment analysis questions are presented for Australian home builders and airlines. However, multi-class classification based on sentiment intensity is not included in these challenges by the authors of [9].

A new approach to sentiment analysis is presented, addressing issues of domain dependency and labeling cost. The two stages of the method are unsupervised ensemble learning and contextual analysis. In both stages, the SentiWordNet sentiment lexicon is utilized. A successful technique is achieved by changing the k-means algorithm, which is the fundamental learning component, and using effective contextual procedures. Experiments on datasets from various domains demonstrate that the approach enhances clustering performance in terms of accuracy, stability, and generalizability. The author also presents novel sentiment analysis challenges. However, the proposed approach does not consider sentiment strength in its multi-class analysis. It instead prioritizes enhancing clustering performance across diverse domains, neglecting to specifically accommodate variations in sentiment intensity within each class in the work from authors [10].

The researchers extensively analyzed academic publications on sentiment analysis from the Scopus database, using quantitative and qualitative analyses. It discussed how sentiment analysis has developed historically, how quickly it has expanded, and how research priorities have changed from

classic areas like product evaluations to more modern ones like social media texts. Classification efforts included Latent Dirichlet Allocation and qualitative coding to offer a comprehensive overview of the field's research landscape. Additionally, top-cited papers were evaluated to provide valuable insights for researchers and contribute to an enhanced understanding of sentiment analysis as an evolving scholarly domain by the authors of [113].

To determine client preferences, the author conducted sentiment analysis on data from customer reviews. They determined the sentiment word strength, examined subjective expressions at the phrase level, and clustered the words into different intensity-based groups. They observed notable differences in their results when contrasting their approach with star-rating systems. In order to provide a clear knowledge of consumer preferences and behavior, they also supplied a visual representation of their data. But they didn't focus on the features of the product to find out how consumers expressed their feelings about it in the work from authors [142].

The authors suggest a semi-supervised method for sentiment classification that makes use of spectral approaches to find reviews that are clear-cut. Then, via a combination of active learning, transductive learning, and ensemble learning, these reviews are utilized to categorize ambiguous ones. Although spectral clustering addresses the shortcomings of k-means clustering, it might not be able to segment reviews well enough. The writers tackle the reviews one at a time to answer this. Five sentiment classification datasets, including the movie review dataset and four datasets with reviews of four distinct Amazon products, are used by them to assess their approach. There are no phases in the procedure especially built for sentiment classification by the authors of [46].

Analyzing sentiments on social media poses challenges due to the large volume of unstructured data and the ever-changing nature of online conversations. Researchers in this field strive to gain insights into public sentiment and behavior by exploring how sentiments are conveyed and interpreted in online communities, contributing to our understanding of digital-age social dynamics and human interactions. It would be beneficial to explore some articles that provide different perspectives on sentiment analysis on social media data.

The researchers used methods like topic modeling, document clustering, and opinion mining algorithms to extract general user opinions and feelings on the themes mentioned during a content analysis of a customer support

platform. For training and inference, a dataset of tweets from English-speaking users addressing the @Uber_Support platform in 2020 was utilized; however, the work was restricted to the Uber brand domain and did not delve into a more thorough examination of user viewpoints in the work from authors [121].

An increase in the quantity and emotions of tweets about face masks was discovered by the author, who examined one million tweets between March and July 2020 to gain insights into how society views COVID-19 and how to prevent it. Automatic text summarization was used to create narratives and group tweets into high-level themes through the use of sentiment analysis, clustering, and natural language processing. In the moment, this method can be useful for evaluating public reaction to health initiatives by the authors of [148].

During the epidemic, the author examined one million tweets between March and July 2020 to gain insight into the public's perceptions of masks. Utilizing sentiment analysis, clustering, and natural language processing, they categorized tweets about mask-wearing into broad themes and provided synopses of the stories behind each subject. Predictive modeling is a tool that health organizations can use to help plan and optimize outreach programs, but they did not identify populations that are not reached by present campaigns in the work from authors [148].

Based on clustering, the author created an architecture for sentiment analysis of social media text data. The three parts of the architecture are the randomized clustering Cuckoo search, similarity discovery, and data cleansing. For a text dataset, the suggested architecture can determine the ideal or nearly ideal number of clusters. The Niek Sanders tweet dataset was used by the authors to test their model, and it was contrasted with six other algorithms, including K-Means, LDA, SMSC, and GLIC. Nevertheless, they haven't examined the suggested algorithm's possibilities by using it on diverse datasets from other industries, like education, health, and business in the authors of [7].

To create a condensed text database in response to a particular topic query, the authors applied sentiment analysis to social media data analysis and explored the application of clustering techniques to the sentiment analysis results in the work from authors [82].

By utilizing emotional signals from social media to model two types of signals—emotion correlation and indication—the authors suggest studying unsupervised sentiment analysis. Their strategy is compared with the most

advanced techniques on two Twitter datasets, and these signals are integrated into an unsupervised learning framework. It is their goal to learn more about how sentiment analysis is affected by emotional cues in the authors of [81].

Semantic analysis and feature extraction play a critical role in sentiment analysis, providing the groundwork for comprehending the significance and environment of written text. Scholars derive significant attributes from the meanings of words and expressions to grasp emotions conveyed in the text. By employing inventive methods, they uncover linguistic subtleties, leading to more precise sentiment analysis. Delving into the literature on semantic analysis and feature extraction in sentiment analysis can provide additional insights into this area of study.

Clustering features is a novel approach to feature extraction that is presented. The goal of the suggested feature extraction method is to cluster semantic features in order to reduce the data simplicity that supervised sentiment analysis encounters. Features for clustering that have been suggested include semantic information and reducing data sparsity for machine learning methods. Support vector machines and Boolean Multinomial Naive Bayes (BMNB) machine learning techniques are utilized for classification in all of the trials. However, no comparison has been made using different approaches to create the feature clusters through experimentation in the work from authors [4].

A novel approach to feature-based clustering and weighting for sentiment analysis on Twitter tweets has been put forward, taking into account word discriminability/dependency and part-of-speech tagging. Emotional terms are prioritized for accuracy in a multinomial Naive Bayes model as well. An efficient sentence classification technique is described for informative data selection, and a Bayes-based text classifier is employed for feature weighting by the authors of [173].

Using a quick clustering method like K-means, grouping similar microblogging tweets can represent significant attitudes about an item. Microblogging, on the other hand, produces a high-dimensional dataset with sparse data because the messages are brief and chaotic. The solution involves choosing pertinent features using the tf-idf technique and reducing the high-dimensional dataset while keeping the most relevant features using the Singular Value Decomposition (SVD) technique. The optimal starting state of centroids is found by applying the artificial bee colony (ABC)

algorithm, which also addresses the convergence of K-means to a local optimum. SentiWordNet is used to examine each group's sentiment polarity after clustering into K groups, with the ideal K being determined by silhouette analysis in the work from authors [129].

A technique based on clustering vocabulary terms using a collection of opinion words from a sentiment lexical dictionary is presented for creating a feature set for sentiment analysis using Word2Vec. On the Internet Movie Review Dataset, the resultant feature set is classified using two classifiers: Support Vector Machine and Logistic Regression. However, more research is required to see how well Word2Vec performs in terms of confusion and in comparison to other lexical dictionaries with more features in the authors of [14].

In healthcare, sentiment analysis provides valuable insights from patient feedback, medical forums, and discussions on health-related topics. Researchers use this technique to understand patient sentiments, identify trends, and assess healthcare service quality. This approach has the potential to improve patient satisfaction, enhance healthcare delivery, and inform decision-making in the medical field. Some articles have explored the application of sentiment analysis in healthcare, discussing its potential benefits and challenges as follows.

Researchers have created a novel method that leverages machine learning and a domain-based knowledge vocabulary to extract semantic relations from medical concepts. The study includes assigning positive and negative attitudes to various medical concepts and situations, such as diseases, symptoms, medications, human anatomy, and other medical words. Features like parts-of-speech, gloss, similar sentiment words, affinity score, gravity score, polarity score, and sentiment are all provided by the lexicon. The system was assessed using supervised classifiers such as Sequential Minimal Optimization based on support vectors, Logistic Regression, and Naive Bayes. The work produced a concept clustering application and may find use in recommendation systems and medical ontologies; nevertheless, it did not make any attempt to improve the system by adding automatically detected patterns from contexts in the work from authors [119].

A framework for analyzing user-generated material in health communities is proposed by researchers. Three steps comprise the framework: first, medical phrases are extracted; second, a weighted scheme called conLDA is added to Latent Dirichlet Allocation (LDA); and third, grouped themes

are analyzed by sentiment polarity and physiological and psychological attitudes. To obtain in-depth understanding, however, they have not included sentiment intensities in their sentiment analysis or carried out cross-analyses of positive, negative, physiological, and psychological emotions by the authors of [181].

The research gathers real-time health updates from reliable websites, where users discuss their side effects and medication experiences. The author's goal is to compile user posts for each medication and offer insightful analysis for patients and the medical community. They also intend to categorize consumers according to their emotional states. Additionally, by employing Association Rule Mining to find drug-related trends in user posts, the study gains knowledge on symptom-drug segment in the work from authors [111].

Sentiment analysis within specific domains, like product reviews, political discussions, educational platforms, and health communities, allows for a deeper understanding of emotions in these specialized contexts. This can provide valuable insights into customer preferences, political opinions, educational outcomes, and healthcare experiences. Customizing sentiment analysis techniques to particular domains enables researchers to uncover detailed sentiments and recognize emerging trends for better-informed decision-making that is customized to the unique needs of each domain. There are articles available that explore domain-specific sentiment analysis and its importance across different fields.

The authors examine various clustering methods for sentiment analysis and propose a method for identifying connections between tweets from political leaders in many nations about the subjectivity and polarity of opinion around the growing use of pharmaceuticals and medications for the treatment of COVID-19 by the authors of [30].

Examining microblogging as a kind of digital word-of-mouth, the writer examined more than 150,000 microblog entries that featured remarks and viewpoints related to branding. They looked at the general arrangement, different expression kinds, and the movement of sentiment while contrasting automatic and manual sentiment classification techniques. Through the use of a case study methodology, they examined the content, timing, frequency, and range of tweets from a corporate account. Although tweets' linguistic structure is similar to that of natural language, brand hijacking on microblogging platforms was not examined in this study in the work from authors [87].

With a particular focus on the extraction of pertinent information on obesity and health from public Twitter data, the author seeks to investigate the possibilities of data mining techniques when applied to social networks. A review of the researcher's project outcomes and an assessment of social networks' applicability for data mining are also included in the study in the authors of [92].

Opinion mining and sentiment analysis are the processes of formalizing the study of opinions and sentiments. A significant amount of opinionated data has been generated by the expansion of the digital world, which scholars are attempting to analyze and assess using various methods in the work from authors [156].

2.1.3 Objectives and scope of the sentiment analysis in medical web pages.

The chapter aims to explore the influence of online information about muscle and bone problems on people's emotions and behavior, particularly focusing on back pain, hip issues, and knee joint problems.

Additionally, it seeks to investigate the factors that contribute to the popularity and trustworthiness of certain web pages compared to others, with a focus on emotional content. To examine these concepts empirically, computer programs are employed for sentiment analysis of web page text.

The sentiment analysis of medical web pages can provide valuable information on patients' emotions, behaviors, and attitudes toward muscle and bone problems such as back pain, hip problems, and knee joint problems.

Understanding emotions on health-related web pages is deemed significant because it could influence the individual's experience of pain and decision-making regarding their health. For example, if a web page evokes strong feelings of concern about pain, it could influence how one copes with that discomfort. In addition, sentiment analysis can help identify and address misconceptions or misinformation that may exist on medical web pages.

The chapter aims to verify the validity of these concepts and address the ethical implications of using emotions to impact people's health-related decisions. It is crucial to note that this research does not directly involve individuals but rather utilizes publicly accessible web data.

In essence, this chapter seeks to explore the influence of online emotions on our perception of healthcare issues and decision-making processes related to our well-being. Additionally, it delves into the ethical considerations

surrounding the use of emotions in disseminating health information.

The purpose of the chapter is to analyze the sentiment expressed in medical websites related to muscle and bone problems, specifically focusing on back pain, hip issues, and knee joint problems. This analysis will provide valuable insights into how emotions on these web pages can impact individuals' experiences of pain and their decision-making regarding their health.

2.2 Methodological approaches in sentiment analysis of medical web pages

In this section, we will outline the procedures we followed to collect and organize the data for our investigation. Initially, we identified crucial search terms related to orthopedic and medical conditions such as back pain, hip replacement, and knee replacement. These terms acted as the foundation of our research.

2.2.1 Sources of medical web page data.

In our study, we conducted an extensive investigation of online resources on orthopedic conditions using SEMrush Competitors Research. This robust digital marketing tool enabled us to carefully examine Web pages dedicated to topics such as back pain and hip and knee prostheses, as well as related medical discussions found on various forums and social media platforms.

Through the use of SEMrush in Figure 2.1., we carefully analyzed patterns and evaluated web pages based on specific criteria related to their content. This thorough procedure resulted in the gathering of around 2000 web pages that were thoughtfully classified according to different conditions, providing a wide-ranging and extensive dataset.

The subsequent phase involved an in-depth analysis using SEMrush to refine and group these pages into condition-specific datasets, laying the groundwork for our study's data pool.

Our subsequent crucial phase entailed exploring the affective substance incorporated in these medical web pages by implementing sentiment analysis methodologies. This procedure necessitated converting unprocessed text from page URLs into organized data, laying the groundwork for our examination of emotional content.

In summary, our methodology focused on thorough data gathering and pre-processing. Utilizing SEMrush for source identification and employing sentiment analysis methods played a pivotal role in extracting an extensive

emotional dataset from web pages. These procedures were vital to guarantee the dependability and pertinence of our dataset for a subsequent detailed examination.

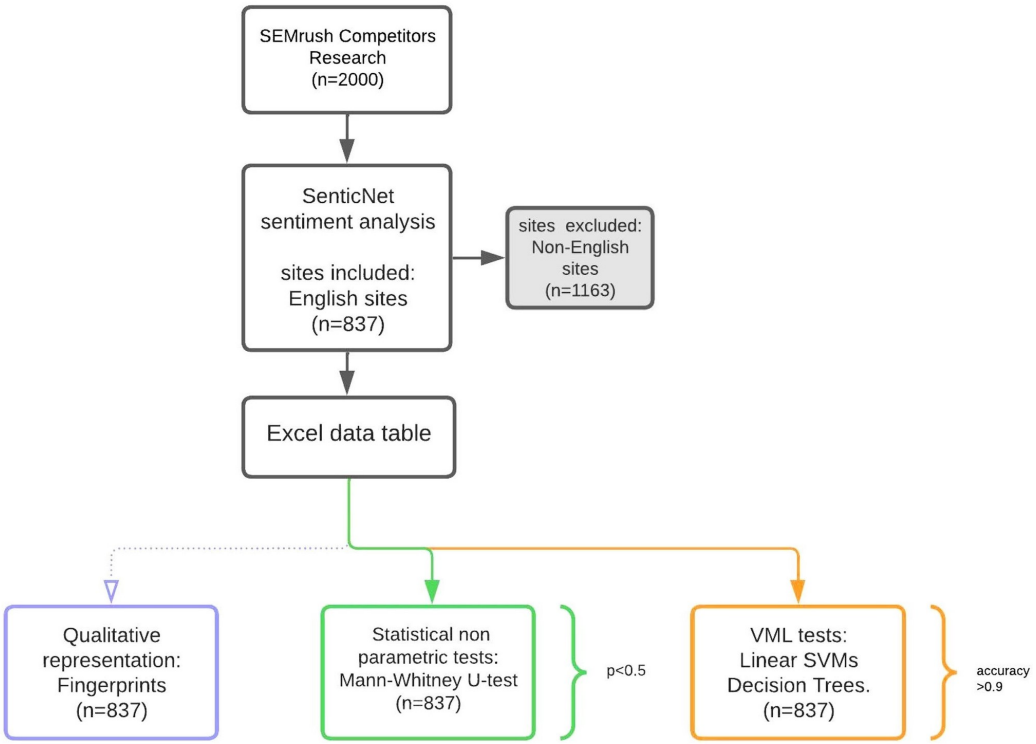


Figure 2.1: Block diagram showing the workflow's steps. Here, SEMrush set the maximum number of sites that might be included in the initial study at 2000 at random. Senticnet retrieved the count for every word pool that was identified as belonging to a single emotion. The fingerprint graphic format was designed to provide a qualitative illustration of the emotional pattern associated with each illness. The variables were subjected to statistical analysis (Mann-Whitney U-test) to determine the significance of the association before being assigned to SVM linear testing.

In the following portion of this section, we will elucidate the key metrics used in the analysis of website performance and Search Engine Optimization (SEO). These metrics include Authority Score, Referring Domain, Backlinks, Search Traffic, and URL Keyword, each providing insights into the quality, relevance, and visibility of a webpage or domain on search engines like Google.

(1) Authority Score (AS): SEMrush utilizes a standardized metric to

assess the overall quality and impact on Search Engine Optimization (SEO) of a URL or domain. This measure accounts for a number of variables, including traffic from organic searches, referring domains, backlinks, and other relevant aspects. It works as a tool for analyzing the relevance of web pages or domain linkages To be clear, a backlink is a reference that functions similarly to a citation and is made from another website (referrer) to an online resource (referent), such as a website, web page, or web directory. A higher Authority Score indicates a greater perceived value assigned to the backlinks associated with a particular domain or webpage.

(2) Referring Domain: Referring domains are websites that link to another website being analyzed for backlinks. Google determines a domain's trust based on its backlink profile and gives importance to has an abundance of referring domains. The total number of referring domains with at least one link pointing to a certain URL is counted by this metric. When taking this measure into account, SEMrush only takes into account the domains that it has seen recently.

(3) Backlinks: Backlinks are links that connect one website to another, and they play a significant role in search engine rankings, particularly for search engines like Google. This metric shows how many backlinks there are overall that point to a certain URL. The backlinks that SEMrush has found in recent months are taken into consideration.

(4) Search Traffic: The total amount of visitors arriving through a particular channel from various sources is referred to as "organic traffic". This indicator shows how much organic traffic, estimated over a certain time period, was routed to a specific URL related to the term under analysis.

(5) URL Keyword: The quantity of terms for which a specific URL appears in search engine results.

2.2.2 Semantic analysis—Senticnet

Website URLs were used to extract raw text data, which was then exported as a CSV file. Numerous types of orthopedic disorders or ailments, such as back pain, hip prosthesis, knee prosthesis, etc., are included in the information gathered. Opinion mining and emotion artificial intelligence were used as sentiment analysis approaches to examine the opinion mining and emotion associated with the data. In this study, SenticNet was utilized for the analysis process (https://sentic.net/). This approach to opinion mining combines multiple disciplines such as semiotics, psychology, linguistics, and machine learning. It is described as the third block in Figure 2.1.

Sentiment computing is a multidisciplinary paradigm that, as opposed to statistical sentiment analysis, places an emphasis on maintaining the semantic representation of linguistic concepts and sentence structure. SenticNet classifies emotions and efficiently expresses the affective information present in natural language text by using the Hourglass of Emotions model in the work from authors [31].

Emotions are categorized into four distinct dimensions under the categorization framework, each of which has varying degrees of activity and influences the total emotional state. Based on their Pleasantness, Attention, Sensitivity, and Attitude, affective states are categorized. The intensity of the emotion experienced is determined by the six "sentic levels" of activation that comprise each dimension.

Emotions like Attitude, Sensitivity, Pleasantness, and Attention are all part of the Aptitude dimension. There are several states of pleasantness, including joy, ecstasy, contemplativeness, sadness, and mourning. While awe, surprise, distraction, interest anticipation, and alertness are present in the Attention dimension, anxiety, fear, apprehension, anger, and rage are found in the Sensitivity dimension.

To perform sentiment analysis, we utilized BabelSenticNet in the work from authors [169], a multilingual knowledge base that operates at the concept level. This system recognizes emotions like joy, admiration, surprise, fear, disgust, anger, sadness, and interest by using SenticNet for emotion recognition. In order to effectively categorize and arrange the documents and extract insightful information from them, natural language processing was used by the authors of [60].

Three procedures have been used to prepare the text.

Tokenization: Tokenization is the process of dissecting a given text into the smallest units that make up a sentence, known as tokens. The output of tokenization, for the sentence "Hip replacement surgery can help relieve", for instance, might be: "Hip", "replacement", "surgery', "can", "help", "relieve".

Tokenization simplifies text analysis and processing by segmenting the sentence into discrete words or tokens in the work from authors [145].

Lemmatization: Lemmatization is the process of figuring out an original word's regular form using a dictionary. This technique is particularly useful in text summarization as it helps to reduce redundancy and streamline the representation of words, ultimately improving the accuracy and efficiency of automatic summarization systems in the authors of [120]. For

instance, taking words down to their simplest form—for example, turning "walking" and "walked" into "walk".

Part of Speech Tagging (POS-Tagging): Labeling words in a text with their respective word types—noun, adjective, adverb, verb, etc.—is known as POS-tagging. This technique converts a sentence into a list of words, or tuples, with each word or pair labeled with the appropriate part of speech.

By utilizing sentiment analysis, a systematic approach was used to identify, extract, quantify, and analyze affective states. In addition to the mood factors, a spreadsheet had additional data, like the quantity of words, phrases, and content words. The documents were arranged according to the conditions associated with each one. Each subset of documents within a specific group was represented by stacking emotional vectors into matrices. In these matrices, emotions corresponded to columns, while documents served as row indices (see in Figure 2.4).

2.2.3 Nonparametric statistical tests and machine learning

We compared English language documents using statistical nonparametric tests to assess univariate emotional score group differences between documents for different health condition and assessed the predictive power of various machine learning binary classifiers, such as Naive Bayes, Multi-Layer Perceptrons, Decision Trees, and eXtreme Gradient Boosting, in terms of which health condition a given document is related to based on the emotional content found within the document. The outcomes demonstrated that these approaches' categorization quality was comparable.

The decision to use only linear Support Vector Machines in our study was based on several key considerations. Linear SVMs are known for their efficiency and scalability, especially when handling high-dimensional data, such as the wide range of features extracted from web pages in our study, including emotional patterns generated through sentiment analysis. They can manage these high-dimensional feature spaces without computational bottlenecks, making them a practical choice. Additionally, linear SVMs are less prone to overfitting compared to more complex models like non-linear SVMs with kernel functions. By opting for a simpler model, we aimed to balance model complexity and generalization performance while ensuring our classifier could effectively generalize to unseen data and minimize the risk of overfitting the training set. Thus, we concentrated on Linear SVMs since they can effectively divide feature space in a linear form in the work

from authors [161], and Decision Trees since they yield interpretable outcomes by the authors of [83].

By using machine learning approaches to document categorization, we were able to find multiple cases where the classification accuracy was higher than 0.90 for different types of illnesses. This implies that each class's emotional content follows unique patterns that make them easy to identify. We determined several statistics, including sensitivity (recall), specificity, precision, and accuracy metrics, from the confusion matrices in order to assess performance. In order to achieve equilibrium between recall and precision, we employed the F1 score, which takes into account both metrics by calculating their harmonic mean as follows:

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Since a simple average penalizes extreme values (a classifier with a precision of 1.0 and a recall of 0.0 has a simple average of 0.5 but an F1 score of 0), the harmonic mean is employed instead. This stage is mentioned in the final graphic block in Figure 2.1

2.2.4 Overview of sentiment analysis methods and algorithms.

Sentiment analysis methods can be approached as supervised classification problems, as unsupervised methods for sentiment identification in the absence of labeled data, or as a labeled dataset used to train a model in the work from authors [35].

In sentiment analysis, the most popular method is to separate the data into three subsets: one with words that are positive, another with words that are mostly negative, and the last group with words that are neutral. With this method, text may be categorized according to whether it contains positive or negative words, giving each piece of data a sentiment score. The overall sentiment of the text or the sentiments of several texts can then be ascertained using this score. It should be remembered, nevertheless, that doing sentiment analysis on inconsistent or incomplete datasets may result in conclusions that are not very accurate in the authors of [67].

In our study, which involved the application of sentiment analysis to web page documents related to various musculoskeletal health conditions, we encountered unique challenges inherent to the nature of the content. These challenges arise from the diverse sources of information, including top-ranking English-language websites specializing in musculoskeletal health.

The documents cover topics such as non-specific degenerative chronic lumbar back pain, herniated discs, and chronic degenerative diseases of the hip and knee requiring prosthesis (hip prosthesis and knee prosthesis).

Furthermore, Sentiment analysis was used in our study to analyze web pages about musculoskeletal health issues. The variety of information sources, which included well-known English-language websites, brought special difficulties, such as colloquial language and layman's terminology for medical ideas.

To appropriately capture and understand feelings in this particular medical setting, these factors need to be carefully taken into account during sentiment analysis in the work from authors [79].

Additionally, The particularities of consumer language used in social media posts must be considered when performing sentiment analysis in the healthcare industry.

This includes slang, abbreviations, misspellings, and the subjective nature of sentiment in medical discussions. Therefore, researchers and analysts must carefully design their studies, including the selection of appropriate keywords to retrieve social media content and making informed decisions about filtering out non layperson-contributed content or retaining special types of data. Sentiment analysis on vast data sets, such as web pages, online forums, online news, online reviews, social media, and web blogs, can offer insightful information about people's beliefs and attitudes around medical issues in the authors of [17].

Through sentiment analysis of various sources, scholars and analysts can acquire an improved understanding of the public's attitude towards healthcare matters, identify trends and patterns in patient experiences, and make informed decisions for improving patient care and healthcare policies. To sum up, sentiment analysis on medical websites gives doctors a quick overview of the sentiments that patients can find when looking up information about medications in the work from authors [77].

This can help healthcare professionals understand the overall sentiment and opinions patients have about certain illnesses, medications, healthcare services, and treatments. It can also aid in identifying any gaps or misconceptions in patient knowledge, as well as potential areas for improvement in healthcare delivery.

Sentiment analysis is a field that uses a variety of methods and strategies to examine and categorize sentiments found in textual data, especially on

medical websites. These techniques and approaches are essential in understanding the emotions, opinions, and experiences of patients and users in the medical domain. One common technique is machine learning, which involves training models on labeled data to automatically classify sentiments. These machine learning models use methods like Recurrent Neural Networks, Naive Bayes, and Support Vector Machines to identify patterns and relationships in the data and forecast the emotion of fresh textual data by the authors of [141].

Another technique is lexicon-based analysis, which relies on pre-defined sentiment dictionaries or lexicons in the work from authors [6].

These lexicons contain a collection of terms together with the corresponding sentiment polarity, such as positive or negative. When evaluating text using a lexicon-based approach, the sentiment score of a given text is calculated by counting the number of positive and negative words in the text and allocating a sentiment score based on the occurrence of these words. Hybrid approaches use machine learning and lexicon-based techniques to improve sentiment analysis accuracy. To get over their shortcomings, these hybrid strategies combine the best aspects of both methods. For example, a hybrid approach can train a model on labeled data using machine learning algorithms, and then enhance the sentiment predictions using lexicon-based analysis in the authors of [152].

In conclusion, sentiment analysis in medical web pages is essential for understanding patient opinions and attitudes toward healthcare-related topics. This analysis aids in monitoring brand reputation, analyzing customer feedback, conducting market research, and comprehending patient experiences. Overall, sentiment analysis in medical web pages provides valuable insights into patient opinions regarding illnesses, medications, healthcare services, and treatments leading to improved patient care and healthcare policies.

2.3 Result and Analysis for sentiment in medical web pages

In this section on sentiment analysis in medical websites, we conduct a thorough investigation. To identify key findings, we look at identifying relevant sources, do semantic analysis, and employ machine learning and nonparametric statistical testing. Overall, our work makes significant contributions and sets the stage for future progress in this vital field.

In this research, our goal is to bring light on the impact of sentiment

analysis in medical websites by delving into various sources for semantic analysis while utilizing statistical tests and machine-learning techniques to uncover key insights.

2.3.1 Identification of relevant sources

Based on the popularity and relevancy of each 2000-strong collection of websites in English, French, German, Italian, and Spanish for our research topic, we chose them carefully. These sites were compiled into an Excel diagram. From this list, we identified 837 English-language sites. We focused on analyzing the qualitative aspects of all languages present while conducting quantitative analysis exclusively on the English pages to maintain consistency in our results. This approach allowed us to exclude variables related to linguistic influence that may be explored in future studies. The websites addressed conditions related to the musculoskeletal system, including low back pain, hip and knee prostheses, herniated discs, and back discomfort. Among the collection of 837 English sites examined, Table 2.1 displays how these different conditions were distributed.

2.3.2 Semantic Analysis

The distribution of content words related to emotions, normalized by the number of emotion words per page, can be seen in Figure 2.2. Figure 2.3 displays the distribution of emotion content terms for each situation and particular emotion. As seen in Figure 2.4, a qualitative analysis identifies a visual pattern known as emotional fingerprints. The scaled counts of emotional content terms per document are represented by these fingerprints, which are histograms with each histogram denoting a different emotion. Eight histograms are present in each fingerprint, organized from top to bottom. Cold colors signify empty bins, and warm colors indicate bins that are heavily occupied.

The Figure 2.4 pertaining to the five health problems under consideration. Eight "pixel" columns in the fingerprint, arranged left to right, represent various emotions such as joy, admiration, and so on. The intensity of each column grows from top to bottom, according to an arbitrary scale.

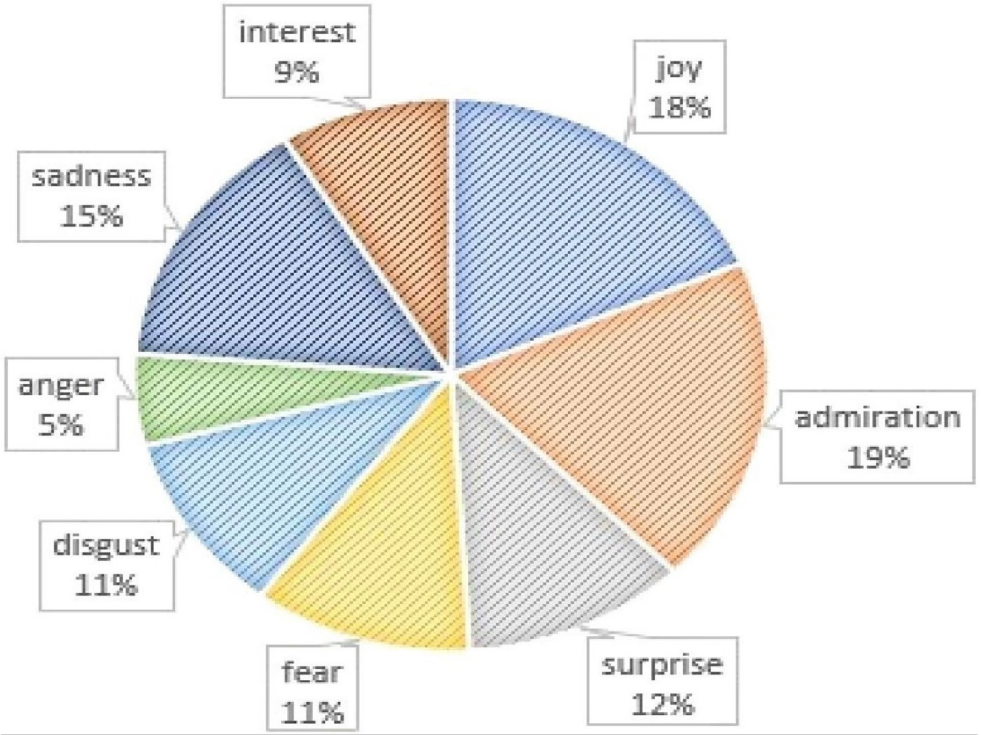


Figure 2.2: Prevalence of emotion words in all selected sites.

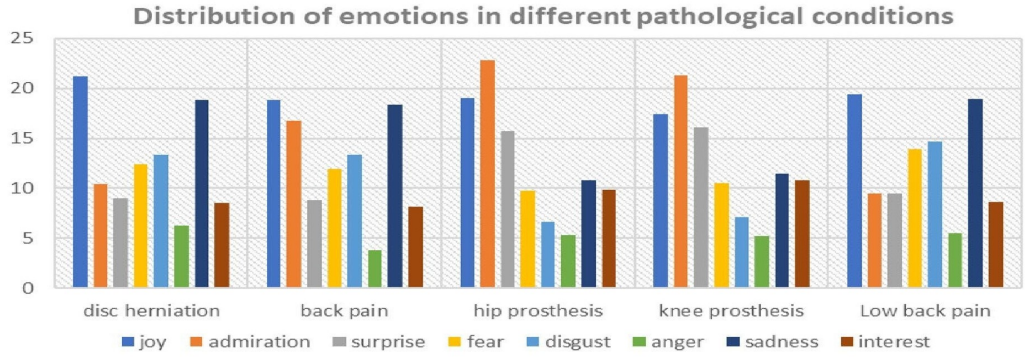


Figure 2.3: Charts showing the frequency of emotion words under various circumstances.

Table 2.1: The number of documents (English sites) considered for each pathology.

Label	Count
Back pain	165
Disk herniation	176
Low back pain	164
Hip prosthesis	168
Knee prosthesis	164

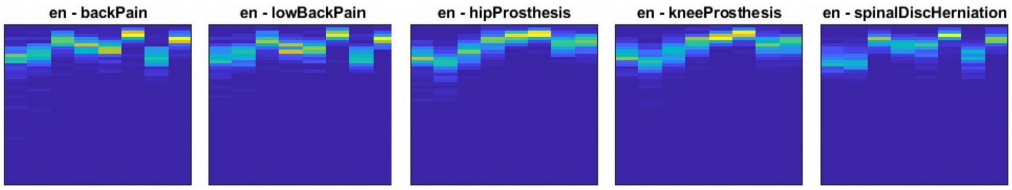


Figure 2.4: Emotional fingerprints of health-related English documents.

2.3.3 Nonparametric statistical tests and machine learning

We chose to contrast the text selections for different health conditions pairwise, employing graphical analysis through scatterplots as well as computational methods such as significance tests and machine learning. Our objective was to ascertain whether there were noteworthy distinctions between the document sets on both an individual and group basis.

In Figure 2.5, the English-language scatterplot displays the distributions of the emotional score variables associated with surprise and disgust for back pain versus hip prosthesis papers.

The digital records were shown as vectors with emotional scores corresponding to various emotions, including joy, admiration, surprise, fear, disgust, anger, sadness, and interest. These vectors were then labeled according to the specific health conditions being studied. Some variables (especially disgust) showed strong discriminatory power to distinguish between different health conditions. Additionally, certain pairs of variables together (like disgust and surprise) also exhibited discriminating capabilities. Some features showed a high correlation with each other; for example, a strong correlation was observed between disgust and sadness. Figure 2.5 illustrates this relationship through a scatterplot comparing the emotions

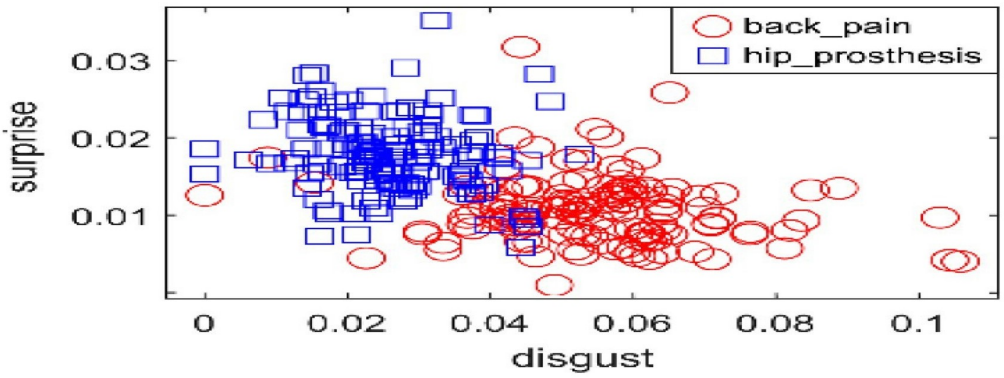


Figure 2.5: Scatterplot of Emotional Scores: Surprise and Disgust in Back Pain vs. Hip Prosthesis Papers

expressed in English language documents discussing back pain versus hip prosthesis issues. The distinct distributions of surprise and disgust in this plot serve as indicators of their significant discriminatory potential.

We intentionally decided not to test Support Vector Machines with non-linear kernels, such as the Radial Basis Function kernel, for several reasons. Our study focused on investigating emotional patterns in web pages related to spine pathology through a binary classification task. We prioritized simplicity and interpretability, which linear SVMs provide by offering straightforward decision boundaries in feature space. Additionally, non-linear kernels introduce complexity that may not be justified given our dataset size and nature, requiring extensive hyperparameter tuning and posing risks of overfitting. Linear SVMs offer significant computational advantages in terms of training time and memory usage, aligning well with our available resources and the exploratory nature of our research. Thus, we found linear SVMs to be the most pragmatic and efficient choice for our classification task.

To assess differences in emotional scores between documents for different health conditions, non-parametric statistical tests were used. In general, nevertheless, these tests did not produce low enough p-values to rule out the null hypothesis that the documents belonging to the two classes are drawn from the same distribution. However, machine learning methods were able to classify a large number of cases with high accuracy, suggesting that the emotional contents of documents show unique patterns for every pair of

medical problems and are thus easily identifiable. All relevant couples' accuracy and other statistics were computed in Table 2.2, demonstrating interesting cases where linear Support Vector Machine classification achieved an accuracy > 0.9.

Table 2.2: Classification Performance of SVM for Health Condition Discrimination. The second to seventh columns contain the confusion matrix [true positives, false positives; false negatives, true negatives], sensitivity, specificity, precision, accuracy, and the F1 score. The linear SVM classifier was trained and validated using a fivefold cross-validation scheme for the English language and for the discrimination between various health conditions in pairs (cases with accuracy > 0.9 were selected). The p-values produced by the Mann-Whitney U-test are displayed in the final column.

SVM	Confusion matrix	Recall	Specificity	Precision	Accuracy	F1 Score	M–W- U
Back pain vs hip prosthesis	$[159, 6; 8, 160]$	0.95	0.96	0.96	0.96	0.96	0.48
Back pain vs knee prosthesis	$[160, 5; 8, 156]$	0.95	0.97	0.97	0.96	0.96	0.049
Disc herniation vs knee prosthesis	$[165, 11; 10, 154]$	0.94	0.93	0.94	0.94	0.94	0.96
Disc herniation vs hip prosthesis	$[162, 14; 11, 157]$	0.94	0.92	0.92	0.93	0.93	0.062
LBP vs hip prosthesis	$[156, 12; 10, 154]$	0.94	0.93	0.93	0.93	0.93	0.2
LBP vs knee prosthesis	$[155, 9; 6, 158]$	0.96	0.95	0.95	0.95	0.95	0.28

The decision not to report the performance metrics of the decision trees system was based on the specific aims and scope of our research, as well as practical considerations regarding result presentation. While decision trees were crucial for data analysis and insight extraction, our focus was primarily

on understanding their interpretability and structure rather than quantifying their performance metrics such as accuracy, precision, recall or F1-score. Introducing these metrics would have complicated the visualization and interpretation process. Additionally, our study followed an exploratory approach aimed at hypothesis generation and insight discovery rather than providing a definitive evaluation of model performance. Thus, prioritizing clarity and interpretability of the decision trees themselves was deemed crucial for understanding emotional patterns related to spine pathology web page success.

Comparable results were obtained when decision trees were used in place of SVMs, and this method allowed for the examination of the impact of emotion variables. The decision tree before and after pruning, a method for condensing the tree by eliminating unnecessary portions that have little effect on classification accuracy, is depicted graphically in Figure 2.6. Pruning reduces classifier complexity, leading to a decrease in overfitting and an improvement in explainability. The predictor importance(have been computed using the decision tree or tree ensemble model used in the machine learning analysis) estimates shown in Figure 2.7 align with the insights of the decision trees.

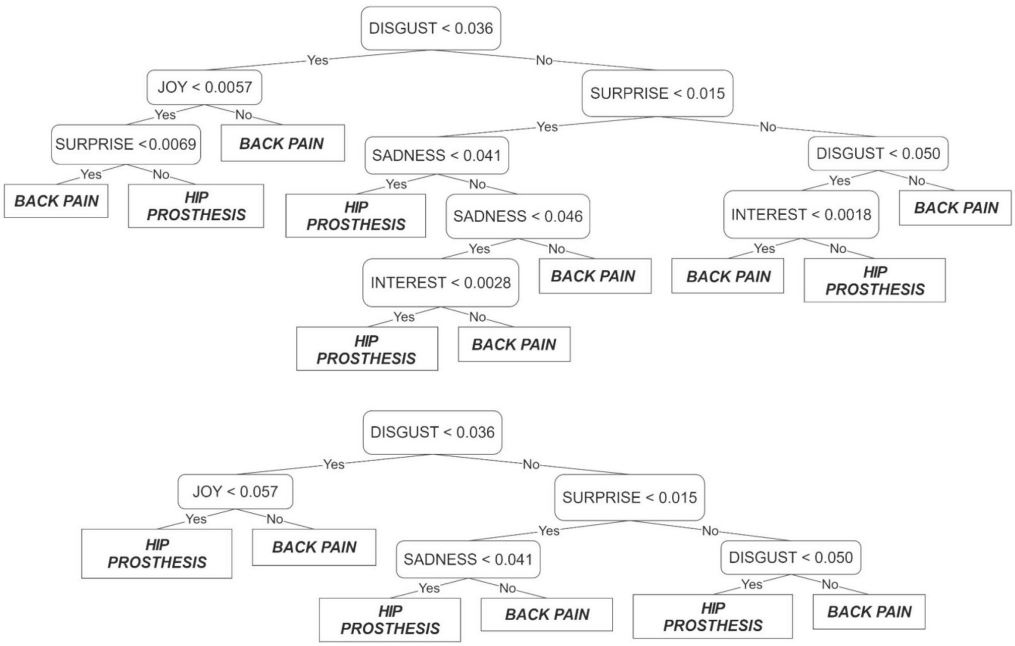


Figure 2.6: Decision tree before and after pruning, for the English language, back pain vs hip prosthesis.

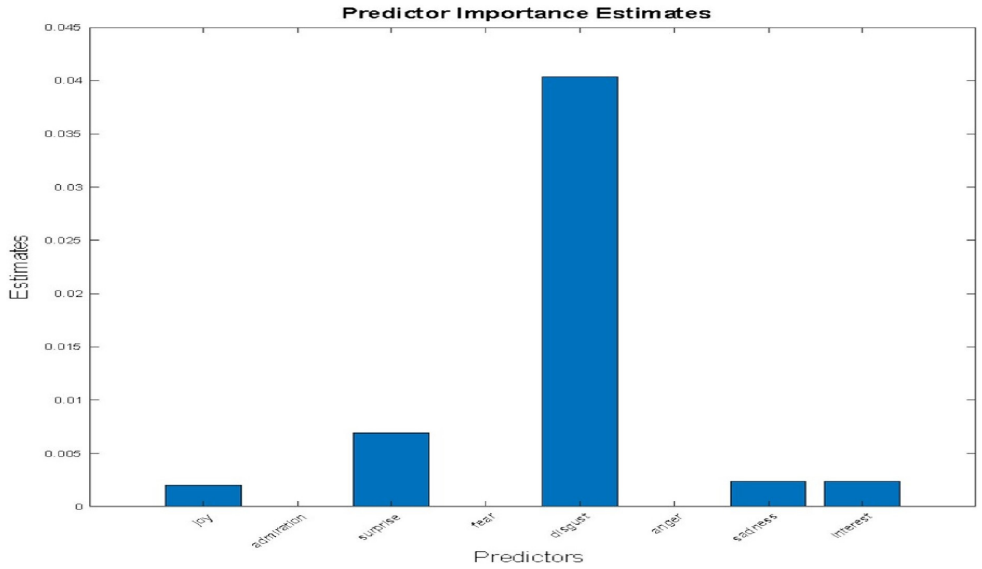


Figure 2.7: Estimate of predictor importance, for English language, back pain vs hip prosthesis.

2.3.4 Discussion of key findings

It is crucial to reframe pain within the biopsychosocial framework and take its emotional component into account in order to obtain a deeper understanding of the experience of pain in musculoskeletal degenerative pathology. Since digitization of healthcare information and increased internet accessibility, there has been a significant change in how health knowledge is accessed. A particular subset of information known as "digital affective collective consciousness" (DACC) is produced as a result of the reciprocal interaction between users and the emotive content of digital medical information. The emotional aspect of the "virtual collective consciousness" seen on popular websites from a variety of fields of human knowledge is incorporated into the DACC. Sentiment analysis is a useful technique, especially when researching circumstances with major psychosocial components, for comprehending how the DACC effects various information sectors or specific themes on the internet. In cases like lower back pain, which often lacks a specific organic cause and instead arises from multiple minor degenerative issues and encompasses emotions similar to fibromyalgia, considering sentiment analysis becomes highly relevant. The somatic marker theory states that emotions play a crucial role in guiding or influencing behavior in the

work from authors [43, 42].

We found that using histograms of scaled emotional word counts per document is a unique and convenient method for visually representing the presence and intensity of emotion-related words.

Through our analysis of digital information, we discovered that disgust contributes significantly to the differentiation of web pages related to low back pain and those associated with hip/knee conditions. Disgust consistently appeared as one of the primary factors considered in ML decision trees generated during our study (see section 2.2);

In certain instances, the level of disgust alone can serve as a pivotal discriminator in delineating the main topic from the emotional patterns extracted from web content related to lower back pain (LBP) and hip/knee affection. This indicates that the intensity of disgust manifests prominently as a distinguishing feature within the affective landscape of these web pages. However, it's noteworthy that machine learning algorithms also incorporate a nuanced understanding of emotional nuances by integrating additional sentiments such as surprise, joy, or sadness. This suggests a more complex interplay of emotional cues within the digital information analyzed, wherein the combination of disgust with other emotions contributes to a comprehensive classification of web page topics. Thus, while disgust emerges as a primary bifurcation point in the decision trees generated by machine learning models, its significance is contextualized within a broader framework that encompasses a spectrum of emotional expressions, ultimately enhancing the accuracy and robustness of the classification process.

The difference in the degree of disgust between the control pathology's emotional patterns and the spine could be a sign of differences in their biopsychosocial profiles. Disgust is a unusual emotion that is not usually connected to being popular. One could argue that digital information about distaste contributes to the adaptive nature of the pain system and encourages avoidance tactics to draw attention away from problems associated with the illness. The success of internet pages may be attributed to a transient decrease in pain sensitivity. It is crucial to keep in mind, nevertheless, that extended exposure to chronic pain might have unfavorable effects.

This raises concerns about the long-term emotional effects of digital information and requires further research with ethical and legal implications.

The role of disgust may vary depending on how different types of affection are perceived. Variations in patient's emotional profiles have been discussed in previous literature reviews. Additionally, both degenerative spinal and

joint conditions affect walking abilities, but through distinct mechanisms in the authors of [157].

The preference for webpages recommended to users based on algorithms indicates a psychological separation between the two sets of conditions that may be influenced by this difference. Unlike other emotions like fear, disgust can be directed towards abstract concepts like moral judgment, hence its structure is less important. This emphasizes the part attraction power plays in the work from authors [146] and which is predicated on the art's inherent aesthetic power; The authors acknowledge some limitations of the current study. As far as they are aware, no research has been published on the evaluation of inherent aesthetic qualities or the function of medical websites, particularly those that discuss disorders of the spine, hip, or knee.

The selection of the control group was made based on a similar medical condition to avoid potential confounding variables but with distinct differences in surgical outcomes. Our results did not demonstrate a statistical association that aligned with predictive power. However, it has been shown that the use of predictive models can enhance explanatory frameworks and establish new theoretical foundations in neuroscience research by the authors of [54].

Our goal in this study was to use machine learning algorithms in conjunction with somatic marker theory to predict disorders based on the emotional content of web pages, surpassing simple statistical connections. In the framework of a digital information society, we also expanded the applications of somatic marker theory to social neuroscience.

This work may have consequences for modernizing the current Biopsychosocial model (bps) as well as for a better understanding of biopsychosocial diseases. A person's health, sickness, and the provision of healthcare are all understood through the lens of the biopsychosocial model, which takes into account biological, psychological, and social aspects as well as their intricate relationships.

Furthermore, we introduced and utilized an innovative fingerprint methodology as a representation tool for emotions, particularly suitable for comparative tests such as intralinguistic experiments.

Further investigations are required to comprehensively delineate and validate the implications of these findings, thereby establishing a solid framework for potential future applications. To reduce any biases related to language reliance, the quantitative analysis was limited to the English language only. Future research on inter-language comparison may focus on

fingerprint analysis (Figure 2.4). Senticnet simply considers the existence of words linked to particular feelings. More sophisticated approaches, however, also take into account additional elements like text complexity, which includes sarcasm or negation, and the valence of emotive words, which varies based on the context and domain. Despite this simplified approach being commonly used in many studies, it may still be effective in practice in the work from authors [147].

In summary, our study found that ML predictivity revealed a distinct pattern of emotions that are strongly associated with specific medical conditions when considering "successful" medical web pages. It was specifically demonstrated that the most basic emotion with a considerable discriminative capacity between degenerative disorders of the spine and great joints was disgust, indicating variations in psychosocial profiles. Numerous explanations have been proposed and will be investigated in more detail in subsequent studies.

A framework for upcoming research is presented, based on the DACC concept. It is advised that additional studies should examine multiple languages for comparison and analyze the emotion of digital material provided by patients through social networks. Furthermore, exploring the relationship between digital emotional content and real-life emotions as well as broadening the focus to include other DACC domains and disorders may result in the development of novel behavioral analysis models. By utilizing this information, it might be able to get over restrictions in therapies that rely on mechanistic pathogenetic reductionism and switch to a more complete BPS model that takes DACC's involvement into account. This process has the potential to improve public knowledge of accurate information, optimize approaches for major health issues, apply neuroscience evidence to medical practice, deepen our understanding of how internet information spreads, and ultimately improve community well-being and healthcare outcomes.

2.4 Conclusion and Future Directions

In the modern digital era, the internet has emerged as a vital source for individuals seeking medical guidance and knowledge. Healthcare websites play a pivotal role in offering information on medical conditions, treatments, and services.

This study explores the relationship between emotional content in online

medical information and the biopsychosocial nature of pain experiences in musculoskeletal degenerative pathologies. By using sentiment analysis and machine learning techniques, we uncover distinct emotional patterns associated with conditions like lower back pain, spine issues, and joint degenerative pathologies. The findings highlight the prevalence and intensity of disgust as a key factor that differentiates spinal-related ailments from joint-related ones, providing insights into significant psychosocial differences. The role of disgust in medical contexts can influence the success of web pages by potentially triggering avoidance strategies and temporarily reducing pain sensitivity.

In the future, it would be valuable to conduct cross-language comparative studies to explore cultural differences in digital emotional content. Additionally, analyzing patient-generated content on social networks could provide insights into real-life emotional alignment.

Expanding our analyses to other sectors of DACC and different pathologies would contribute to more robust behavioral models. Integrating findings from DACC into the biopsychosocial model has the potential to enhance healthcare strategies while incorporating neuroscience and machine learning insights can refine medical practices and address important health issues.

Ultimately, this investigation of DACC provides a profound understanding of the complex relationship between digital emotional content and musculoskeletal pathologies, leading toward holistic approaches that consider how digital information affects patient experiences and behaviors in healthcare settings.

2.5 Our Significant Contributions for this chapter

This chapter presents an innovative investigation into the analysis of emotions and opinions, focusing specifically on musculoskeletal degenerative pathology and digital medical information. Through our study, we have made remarkable contributions that greatly enhance this area.

Unveiling Emotional Content: This study represents a significant milestone in the investigation of emotional aspects within digital medical information, particularly concerning web pages related to musculoskeletal degenerative diseases like lower back pain and hip/knee conditions. By analyzing how emotions are incorporated into the dynamic interaction between users and digital content, we have shed light on the psychosocial components of experiencing pain.

Identifying Disgust as a Key Discriminant: Our findings suggest that the emotion of disgust plays a significant role in differentiating web pages associated with low back pain and hip/knee conditions. Through machine learning decision trees, we consistently observed the inclusion of this emotion, indicating its importance in distinguishing emotional patterns between these two conditions.

Pioneering Emotional Fingerprints: We have developed an innovative approach called "emotional fingerprints" that visually represents the prevalence and intensity of emotion-related words in web pages. These histograms, which display scaled emotional word counts per document, provide a quantitative view of emotional content within digital medical information. This novel method offers a unique and valuable way to analyze emotions in online healthcare resources.

Application of Sentiment Analysis Techniques: We examined affective states in our dataset using sentiment analysis techniques. We analyzed emotions quantitatively using text analysis, natural language processing, and computational linguistics techniques. Employing machine learning algorithms allowed us to accurately predict the original topics of web pages based on emotional patterns. This highlights the practical applications of sentiment analysis in healthcare-related information.

Implications for Healthcare Providers: The implications of our findings extend beyond academic settings and have important consequences for healthcare information providers, institutions, and policymakers. It is crucial to recognize the ethical considerations involved in understanding the impact of emotional elements on medical information that is distributed online. This highlights the significant but often overlooked role emotions play in web content.

We examine sentiment analysis in musculoskeletal diseases and its consequences for digital health records in this chapter. In this context, we obtain important insights on the impact of emotions on patient experiences and behavior.

2.6 My personal contributions for the Sentiment analysis in medical web pages

Throughout our research journey, I took a hands-on approach to uncovering the emotions and opinions expressed in digital health information concerning muscle and joint issues. I played a key role in gathering relevant text

from various medical websites through textual data scraping, ensuring our dataset was comprehensive and diverse.

I initially undertook several important actions when we first started exploring how people express their emotions online. Initially, I organized and simplified the text, divided it into smaller segments, and tried to understand the true meanings of words. Before delving into our main research study, I reviewed numerous other scholarly articles to ascertain existing knowledge on the subject.

Before finalizing our findings, I conducted a preliminary statistical analysis to understand the data trends and patterns. This initial exploration helped me identify potential areas of interest and guided our subsequent in-depth investigation. Additionally, I performed sentiment analysis to scrutinize the emotional tone of the content we were studying.

Collaborating closely with coauthors, I shared my preliminary statistical analyses and sentiment analysis to seek their valuable input and feedback. This collaborative effort was essential in refining our interpretations and ensuring the accuracy of our conclusions.

Ultimately, comprehensive involvement in the project significantly enriched the depth and reliability of our research on the emotional dimensions of digital health information.

Chapter 3

Sentiment-Preserving Extractive Summarization

3.1 Introduction

Sentiment analysis is useful for analyzing natural language, particularly in text summarization. Extractive text summarization successfully preserves sentiment and uses existing topic-based technologies. It focuses on locating and extracting the linguistic elements that support the text's emotional content to calculate a sentiment measurement. We shall examine sentiment-preserving extractive summarizing approaches in this chapter. We will talk about how these strategies are applied to summarize a text's substance while preserving the sentiment it expresses.

3.1.1 Definition and significance

The practice of extracting sentiment from a text to classify the tone or viewpoint expressed therein using a model of the hourglass of emotions—such as joy, admiration, surprise, fear, happy, anger, sadness, and interest—is known as sentiment analysis.

Analyzing the emotions represented in the content is one possible technique to look at this aspect. Sentiment analysis is a valuable tool for interpreting user sentiments and opinions that are conveyed in a variety of unstructured text formats, including emails and posts on social media. It is employed in many different fields, such as user preference finding, customer support, and market research in the work from authors [186].

Recently, deep neural networks have been used to apply sentiment analysis to visual data as well. A method used in natural language processing

called sentiment analysis, or opinion mining, is used to identify and categorize the feelings and viewpoints represented in a document by the authors of [22].

Sentiment analysis is of significant importance because it provides businesses with a better insight into client comments and emotional responses. By monitoring online conversations on the Web and social media applications, businesses can gather valuable insights about their brand and products. This information can be used to make informed decisions about marketing strategies and product development, ultimately leading to improved customer satisfaction and loyalty in the work from authors [80].

In our approach to summarization, we have innovatively developed a method known as sentiment-preservation extractive summarization. This technique is designed to condense and extract key points from a piece of writing while consciously maintaining the sentiment expressed in the source material. This way, we guarantee that not only the essential information is preserved, but also the emotional tone and viewpoints are safeguarded in the condensed version. This unique approach reflects our originality in devising strategies for effective summarization.

By analyzing expressed emotions in content, sentiment-preserving extractive summarization creates a concise representation of text while capturing authorial feelings and opinions. It condenses text without compromising its original emotional expressions and viewpoints by the authors of [21].

Therefore, sentiment-preserving extractive summarization is an essential tool for businesses and researchers who want to gain a brief interpretation of a text's essential ideas while also capturing the emotional tone or opinion behind it.

Extractive summarization of the feelings of the customers is especially valuable in domains where understanding the emotional tone and opinions expressed in text data is crucial, such as customer reviews, market research, social media monitoring, and public health. By utilizing sentiment-preserving extractive summarization techniques, businesses can quickly and efficiently analyze enormous amounts of textual data to comprehend the overall sentiment and key points expressed by customers, users, or the general public. This information can then be used to boost customer satisfaction, brand reputation, and decisions based on data. In the context of sentiment-based summarization, the use of existing topic-based technologies is a common approach in the work from authors [45].

In conclusion, extractive summarization that preserves sentiment is a

vital tool for businesses and researchers in diverse fields such as news, scientific articles, stories, etc. It provides a concise representation of the text while capturing the emotional tone or opinion of the author, enabling a comprehensive understanding of the main points within a text. This method is particularly crucial in domains where understanding the emotional tone and the expressed opinions in textual data is essential, such as customer reviews, market research, social media monitoring, and public health.

3.1.2 Methods and algorithms

Various methods and algorithms can be used for extractive summarization that preserves sentiments. Graph-based, machine-learning, and frequency-based methods are frequently used to determine the most significant sentences in a text while maintaining their original emotion. Frequency-based algorithms analyze word and phrase frequencies to identify sentiment-carrying sentences. Graph-based algorithms use graphs to represent sentence relationships and determine importance based on connections and expressed sentiment. Machine learning utilizes supervised or unsupervised techniques to predict sentence sentiment accurately. In order to inform the choice of sentiment-preserving sentences for the summary, topic techniques now in use can also help identify the primary themes or topics in the text by the authors of[41].

By leveraging existing topic-based technologies, sentiment-preserving extractive summarization systems can effectively identify and extract sentences that capture both the important themes or topics of the text and the sentiment expressed. Then, by combining these lines, you can produce a summary that succinctly conveys the key ideas while faithfully capturing the feelings stated in the original text. In conclusion, sentiment-preserving extractive summarization is a valuable technique for accurately capturing the main points and sentiments expressed in textual data.

Extractive summarization can preserve sentiment by using methods and algorithms such as frequency-based, graph-based, and machine-learning approaches. These techniques help extract sentences that convey important themes while maintaining the author's sentiment.

The aim of sentiment-preserving extractive summary is to identify and take pertinent lines from a text while retaining the emotions associated in the work from authors [178].

Sentiment analysis techniques, including frequency-based and graph-based

algorithms as well as machine learning approaches, enable accurate identification and extraction of sentences conveying the original text's sentiment. The original text's sentiment can be preserved in a succinct summary that is produced by arranging and combining these sentences by the authors of [45].

To achieve sentiment-preserving extractive summarization, existing topic-based technologies can be leveraged to figure out the main focus or topics in the text. These themes or topics can then guide the selection of sentiment-preserving sentences for the summary, ensuring that both the main points and sentiments of the original text are effectively captured.

Sentiment-preserving extractive summarization is a crucial tool for distilling important aspects of a text while retaining the emotional nuances conveyed by the author. By combining sentiment analysis techniques with topic-extraction technology, the process accurately captures key themes, topics, and underlying sentiments expressed in the text. This makes it possible for readers to grasp the content's sentiment and major concepts of the content thoroughly without having to read the full text.

The use of methods and algorithms, including frequency-based and graph-based algorithms, as well as machine learning approaches, creates a strong foundation for extractive summarization while preserving the sentiment. These techniques enable the extraction of sentences that capture important themes and maintain the author's expressed sentiment. In today's rapidly changing world, the significance of accurate sentiment-preserving extractive summarization cannot be overstated.

3.2 Literature on Existing Extractive Summarization

In this comprehensive literature review, we extensively explore various areas of extractive summarization research and categorize existing studies for in-depth analysis. We cover methodology-based articles, domain-specific articles, evaluation-based articles, feature selection and clustering articles, and improvement focus articles. This extensive overview offers valuable insights for further research in the field of extractive summarization.

In the field of methodology-based articles, scholars have examined numerous creative approaches to improve the extractive summarization process.

In extractive summarization, Unsupervised learning methods autonomously extract key insights from text data using techniques such as clustering and

topic modeling, allowing for scalable and adaptable summarization without the need for labeled examples. In the following, we have presented some articles that are related to unsupervised learning methods.

The author proposes an unsupervised text summarization method that uses multi-round calculations to reduce redundancy and include key content by optimizing sentence relationships. Their technique is tested on Chinese, English, and other language datasets, but they did not make any advancements in automatically identifying phrases or differentiating between primary and secondary content in the work from authors [164].

The authors provide an unsupervised approach to multi-document summarization using BERT sentence embedding. They test the model on the DUC 2002-2004 data sets and refine it on supervised intermediate tasks from the GLUE benchmark. However, they do not explore its efficiency on other summarization tasks or its potential for language generation by the authors of [96].

The centroid technique and sentence embedding representations serve as the foundation for a proposed unsupervised extraction method for multi-document summarization. The authors enhanced sentence scoring by combining relevance, novelty, and position in the work from authors [185].

The literature on hybrid summarization methods explores the combination of strategies and algorithms to create comprehensive summaries. This approach utilizes graph-based techniques and machine-learning models to develop adaptable frameworks for summarization.

The author suggests a hybrid method for extractive text summarization that incorporates sentiment analysis, sentence connectivity, and sentence ranking based on key phrases. However, they have not concentrated on enhancing the hybrid technique for higher-quality automatic summaries. The BBC News Summary dataset is used to evaluate the methodology by the authors of [18].

In Bias Reduction methodologies, debiasing algorithms and fairness-aware learning are employed to reduce biases in summarization systems, thereby promoting equity and transparency in content selection.

To address the problem of facet bias in unsupervised summarization models, researchers have developed Facet-Aware Rank, a novel algorithm that uses facet-aware centrality-based ranking. The model employs a sentence-document weight to emphasize various aspects and reduce redundancy. The evaluation is carried out using several benchmark datasets in the work from authors [103].

Graph-based methods for summarization utilize principles from graph theory to represent and analyze textual relationships. By modeling text as a network of nodes and edges, these approaches can extract important information and generate concise summaries that maintain semantic coherence and context. This facilitates more effective knowledge extraction and dissemination by providing valuable insights into the interconnected of concepts and themes within documents. Through the study of several articles, we have sought insights into the utilization of graph-based methods in extractive summarization.

A graph-based summarization method is introduced, using topic modeling and semantic measurements to enhance similarity computations. The concept was evaluated against state-of-the-art methods using datasets from CNN/Daily Mail and Opinion; however, in the work from authors [26] did not expand its application to graph clustering-based summarization methods.

A graph-based method for generating concise and informative summaries through analyzing textual connections created by the authors. They tested the approach using data from the DUC2004 summary task, showing its potential to create excellent summaries without requiring discourse analysis. While the method emphasizes syntactic links for coherence evaluation, it does not address semantic relationships between elements in this improvement effort in the authors of [172].

Modifying the PageRank algorithm with normalized bigram counts, researchers use a graph-based method for Hausa text summarization. It outperforms baselines on a Hausa news article dataset, improving performance in Rouge metrics. However, it lacks an extension to multi-document summarization and reduces redundancy in the work from authors [27].

Documents are modeled as weighted graphs for key phrase extraction and summarization by the authors, with sentence link priors enhancing clustering quality. Using the mutual reinforcement principle, saliency scores are computed for significant phrases and sentences. These values are then used to rank the phrases and sentences for inclusion in the document's top key phrase list and summaries. This approach allows the creation of a hierarchy of summaries with varying levels of granularity through translingual summarizing techniques in the authors of [184].

The author proposes a new method for summarizing text documents by identifying and removing nodes in the independent set of a graph representation. Their approach, called KUSH, preserves semantic cohesion between

sentences in the summary. Using ROUGE assessment measures, they evaluated the effectiveness of this strategy on the Document Understanding Conference (DUC-2002 and DUC-2004) datasets in the work from authors [167].

The author creates two graph-based methods that include several significant metrics, including identity similarity, topic signature similarity, tf-idf cosine similarity, and Jaccard similarity. The ROUGE assessment toolbox and the DUC 2003 and 2004 datasets are used to assess the methodology. But more sophisticated approaches, such as machine learning techniques, are not tested when it comes to combining scores in the authors of [15].

A Bayesian Optimization-based method for Textrank optimization for legal document summarizing is presented by researchers. The Bill Sum dataset evaluation reveals enhanced performance over baseline models. The importance of hyperparameter adjustment is emphasized, and a BO-based approach is suggested in the authors of [85].

TextRank for integrating online news items, resulting in a multi-document summary with repetitive sentences are used by the authors. In the work from authors [75] employed Maximal Marginal Relevance to generate a shorter and more varied summary but did not address irregular words or initials for improved accuracy.

The authors suggest utilizing inverse sentence frequency-cosine similarity in a modified graph-based text summarizing method and a modified TextRank algorithm. The system shows promising results on news articles, but the authors have not compared it with other techniques or evaluated its performance in detail in the authors of [109].

A method for summarization based on clustering sentences in three stages is presented by the authors, illustrating how summarization relies on both sentence characteristics and similarity measures. Experiments on the DUC 2003 dataset highlight the effectiveness of their approach compared to other options in the work from authors [188].

The researcher presents a method for summarizing similarities and variations between connected papers with a graph illustration. The algorithm utilizes spreading activation to discover semantically related nodes, matching the activated graphs of each document to create a graph representing similarities and differences, rendered in natural language. However, they have not utilized alpha links or systematically extracted semantic distance measures from WordNet, nor have they used both text and thesaurus concepts to link extracts into abstracts by the authors of [112].

Survey Analysis systematically evaluates existing summarization techniques, identifies trends, and guides future advancements by synthesizing findings from various studies and experiments.

In this study, the authors present a thorough analysis of graphical-based extractive text summarizing methods, both supervised and unsupervised. It critically evaluates previous surveys in a tabular format, introduces a taxonomy for summary evaluation measures, and describes various existing approaches while comparing their strengths with concentrate on the DUC2001 and DUC2002 datasets' ROUGE scores. However, it does not cover abstract text summarization methods or discuss research challenges in this area in the work from authors [179].

Methodologies for Semantic Strategy combine machine learning and linguistic principles to generate comprehensive summaries with contextual depth. This process utilizes semantic representations and sentiment analysis to produce coherent and informative outcomes.

The author offers a semantic strategy that integrates machine learning, statistics, and graph-based techniques to create an extractive multi-document summarizer. Using word2vec, the approach learns the semantic representation of words from a collection of documents. Tests carried out on the DUC2002 and DUC2006 datasets demonstrate how successful the suggested plan is. But when creating summaries, it doesn't take sentence placement or summary readability into account by the authors of [28].

Domain-specific summarization research aims to customize summarization techniques for specific fields, taking into account the unique challenges and requirements within those areas. These studies seek to improve the relevance, accuracy, and usability of extractive summarization systems in targeted domains by considering domain-specific characteristics such as vocabulary, structure, and content in the following article.

In the work from authors [52] conducted research on a dataset of over 470k medical documents and their summaries using a BART-based summarization system. They work with free text and structured forms for input and target summaries, as well as adapt a proposed metric to evaluate their system. The data set is derived from the Semantic Scholar literature corpus but lacks improvements in summary targets or connections to external structured data sources.

Research focused on evaluating extractive summarization investigates the methods and criteria used to assess the quality and performance of summarization systems. This research contributes to the advancement of

summarization technology by establishing thorough evaluation frameworks and standards for comparing different algorithms and techniques explained in the following articles. Through standardized evaluation procedures, researchers aim to ensure the accuracy and consistency of summarization outcomes, thus driving meaningful advancements in this area.

The author uses extractive text summary techniques to produce a concise summary of clinical trial material. ROUGE metrics and human assessments are used to determine how effective the strategy is. However, in the work from authors [74] did not focus on addressing specific challenges related to condensing clinical trial reports.

The authors assess the effectiveness of the TextRank algorithm in the evaluation of Indonesian smartphones, taking into account various data conditions such as stop words and typos. They determine scores on Rouge-1 using professional summaries as benchmarks. It is significant to remember that the evaluation is centered on TextRank's performance while evaluating Indonesian smartphones. However, the exploration of alternative algorithms and pre-processing methods is not covered in this study by the authors of [138].

The researchers analyzed 8 extractive summarization algorithms for microblogs in emergencies and found significant discrepancies. More advanced algorithms are needed for effective microblog summarization after a disaster, but the work from authors [55] did not explore combining multiple algorithm outputs to enhance results.

Research on feature selection and clustering for extractive summarization seeks to identify important features and group similar content to produce concise and informative summaries. Various techniques, including measures of sentence importance, analysis of term frequency, and clustering algorithms, are explored in these studies. The objective is to improve the accuracy and effectiveness of extractive summarization systems. This research is crucial to improving the quality and relevance of automatically generated summaries in diverse domains and applications.

The researchers have proposed a novel approach to the selection of features for the nearest-neighbor classifier, which summarizes training texts using the importance measure of sentences. The approach involves two measures for sentence similarity: frequency of terms and similarity to other sentences. These metrics are used to rank every sentence, with the top-ranked sentences being chosen for summarizing. It uses term frequency and

sentence similarity but lacks implementation of multi-document summarization or hidden Markov model methods for unigram, bigram and n-gram features, as well as categorically focused document summarization methods in the authors of [25].

To achieve extractive summarization that preserves sentiment, researchers have applied standard machine learning classification techniques for sentiment classification. These techniques involve classifying the sentiment related to the opinion of a document into categories such as positive or negative in the work from authors [16].

The authors propose a novel multidocument summarization method with sentence overlapping. They preprocess the documents, calculate the features for each sentence, and assign scores based on a learned model to select important sentences with less redundancy. However, in the authors of [136] do not use new algorithms or features to extract important sentences.

The authors provide a novel method for multi-document summarizing based on the AdaBoost machine learning meta-learner algorithm. In the work from authors [117] experiment with 450 news pieces downloaded from various medical websites.

By experimenting with various parameter settings and assessing the results using ROUGE, precision, and recall, the author investigated multi-document summarization methods utilizing clustering. According to their major research, the DUC-2002 dataset's highest ROUGE scores are obtained when sentences that are similar to the centroid of all associated document sentences are used. However, the impact of the techniques on different languages was not compared, nor did they explore more fine-tuned clustering applications for further improvement. Additionally, no experiments were conducted on language-specific features in the work from authors [117, 57].

A novel approach to incremental clustering and document summarizing has been created by the authors. This approach reorganizes sentence clusters in real time, showing hierarchical links as soon as a new document is received. It has been effective on real-world disaster management data and TAC benchmark data in the authors of [171].

Integrating sentiment analysis into extractive summarization methods improves the ability to capture both factual information and the underlying emotions and opinions expressed in the text. This integration allows for the creation of summaries that convey the sentiment and tone of the original content. Scholars are investigating different approaches to smoothly incorporate sentiment analysis into summarization processes, allowing for more

sophisticated and contextually relevant summaries in the following studies.

Addressing challenges in product feature extraction and opinion orientation is crucial to achieving sentiment-preserving extractive summarization for consumer reviews identification in the work from authors [159, 176].

To address the challenges posed by the use of different languages in consumer reviews, researchers have developed unsupervised sentiment analysis techniques that automatically extract product features and determine the sentiments expressed towards in the authors of [180].

The identification of specific product features expressed in consumer reviews is crucial for the preservation of sentiment in extractive summarization. Existing topic-based technologies can be leveraged to extract key features and opinions while maintaining emotional coloring. The seamless integration of sentiment analysis techniques with extractive summarization processes offers valuable insights for decision-making, preserving nuanced sentiments and opinions. In the work from authors [45] acknowledge the connection between sentiment analysis and extractive summarization.

The progress in sentiment analysis and the use of topic-based technologies has greatly improved summarization methods. Preserving sentiments is vital for automated sentiment-based summarization, especially with the growing number of consumer reviews. Feature extraction is crucial for identifying and preserving specific product features and related sentiments expressed in reviews by the authors of [176].

As a result, it naturally becomes desirable and necessary to have a sentiment analysis technique that is both efficient and effective and that can summarize the opinions of customers regarding particular product characteristics. To sum up, the challenge of sentiment-preserving extractive summarization entails the extraction of product attributes from customer evaluations and the assessment of their sentiment polarity in the work from authors [176].

Some researchers have tried to introduce sense disambiguation in the analysis process, which has shown higher precision and lower recall in extracting opinion expressions. The introduction of sentiment analysis techniques in extractive summarization allows for the identification and extraction of sentences that not only capture the important themes and topics of the original text but also maintain the sentiment expressed by in the authors of [45].

Combining various summarization techniques can result in more comprehensive and accurate summaries. By utilizing different methods such as

graph-based approaches, unsupervised learning, and semantic strategies, a comprehensive method can result in more encompassing and diverse summaries. In the following article, we aim to elucidate the integration of various summarization techniques.

The authors propose an efficient approach by integrating the BM25 + and TextRank algorithms, demonstrating its superiority over the baseline methods that use the ROUGE, F1 score, recall, and precision measurements. The authors use an innovative and efficient summarizing technique to investigate robust page-rank algorithms in the work from authors [73].

Through this article, the author presents an improved approach to text summarization by combining BART and TextRank algorithms. The method conducts multiple rounds of summarization, first using TextRank to identify key sentences and then generating a summary with the BART model. Investigations using the CNN/Daily Mail dataset reveal that this method outperforms individual BART and TextRank models, leading to higher average recall rates on ROUGE metrics for more aligned summaries with the main ideas and content of the original texts in the authors of [38].

The author developed a more effective TextRank-based text summarization method for WeChat that takes into account user requirements and sentence characteristics. By incorporating the Word2Vec model, increased the accuracy of the summarization. However, their summarization method has not yet been integrated into the WeChat platform, which could improve knowledge concentration and improve the reading experience for users in the work from authors [39].

To improve the quality of GitHub release notes, an automated method utilizing the TextRank algorithm outperformed the LSA approach when tested on a dataset of 1,213 release notes. However, the authors of [128] did not incorporate additional software artifacts or a sequence-to-sequence model to further improve their quality.

Extractive summarization plays a crucial role in tasks such as document annotation and event extraction by identifying specific information from extensive text material. Researchers are actively developing methods to automatically annotate documents and extract pertinent events, entities, and associations. These methodologies facilitate effective information retrieval and knowledge extraction for purposes like organizing information, conducting search operations, and performing analysis.

The researchers introduce a framework for summarizing multiple documents that includes preparing the documents for annotation, extracting

events, times, and event-time relationships. Sentences are prioritized using scores and redundant ones removed to create a summary. The ROUGE metric is used to assess the framework on the DUC 2006 and DUC 2007 datasets; however, it has not been applied to data from the medical or legal domains in the work from authors [110].

The rise in online product reviews has created a demand for specialized summarization methods that can efficiently extract key information from these reviews. Experts are working on techniques to identify crucial features, sentiments, and opinions to provide valuable insights for consumers, businesses, and decision-makers. This helps evaluate products and facilitate decision-making by condensing lengthy reviews into concise summaries.

The authors developed a technique to summarize product reviews by grouping them based on sentiment and reasoning. They combined several pre-processing techniques., including part-of-speech tagging(POS-Tagging), stemming, and syntactic analysis to identify relevant facets. The study did not address the creation of a module for recognizing and processing proper names by the authors of [108].

The author summarizes Amazon movie reviews using four advanced algorithms and a feature selection technique. Sentiment analysis categorizes the reviews, and a hierarchical summarization method condenses lengthy ones. The experiments are conducted on a large Amazon review dataset, but in the authors of [144] have not explored deep learning integration for improved review summarization.

Condensing legal materials is crucial for summarizing court rulings, agreements, and laws. Scholars are creating algorithms to automatically condense legal writings by identifying main points of contention, judgments, and prior cases. These methods enhance lawful exploration and decision-making by offering concise digests of complex legal documents.

The author describes a technique for creating a summary of each court decision by grouping them according to subjects obtained from Hierarchical Latent Dirichlet Allocation (HLDA) using the same topics. Their technique improves upon previous methods by creating a topic-based model that can categorize legal judgments into distinct clusters and generate a summary for each legal decision within the cluster in the work from authors [168].

Academic researchers are developing algorithms to recognize common patterns, extract vital data, and combine information from various document sets. These methods help with tasks such as literature review, news analysis, and synthesizing information by enabling users to efficiently grasp

important aspects across multiple documents.

The paper suggests a novel approach to multi-document summarizing that makes use of sentence compression and textual entailment relations. It uses an expanded tf-idf approach to rank phrases, formulates the problem as a knapsack problem, and computes the entailment scores. Experiments are carried out on DUC 2007 and MultiLingPilot 2011 datasets. However, in the authors of [127] did not reinforce the ranking strategy with features such as sentence position or sequence-based/tree-based sentence compression.

The textual-unit similarity framework measures the similarity between sentences, paragraphs, or documents to identify relevant content for summarization. Researchers are developing metrics and techniques to group similar units and extract representative summaries. These frameworks generate concise and coherent summaries capturing essential information across text documents.

The author suggests a novel submodular framework based on word coverage and textual-unit similarity for multi-document summarizing tasks. They demonstrate that different variations of summarization tasks can be modeled from the proposed framework. Benchmark dataset experiments demonstrate the accuracy of their methodology in the work from authors [98].

3.3 Introduction to TextRank algorithm

In this section, we are going to explain Page rank algorithm, TextRank algorithm, BERT sentence Embedding, and Cosine similarity. All of these terms are related with generating extractive summarization which are discussed in th following sections.

3.3.1 Page Rank Algorithm

Originally created by Google, the PageRank algorithm is a graph-based ranking system in the work from authors [165]. Every webpage is given a number score according to the relevance of the inbound links on it. The PageRank algorithm uses the quantity and quality of inbound links to determine the relevance or value of websites by the authors of [130].

The PageRank algorithm can be defined as a method used by search engines, particularly Google, to determine the ranking of web pages in search engine results. The PageRank algorithm consists of the following steps:

1. **Parsing and Extracting Links:** The search engine parses the webpages and extracts the outgoing links from each webpage.

2. **Calculating Initial PageRank Values:** Each webpage is assigned an initial PageRank value, typically equal for all webpages.

3. **Calculating PageRank Scores:** Each webpage's relevance is iteratively determined by the PageRank algorithm based on the inbound links that it receives in the work from authors [70].

4. **Updating PageRank Values:** In each iteration, the PageRank algorithm updates the PageRank values of webpages according to how significant the inbound links are.

5. **Convergence:** The PageRank algorithm continues iterating and updating the PageRank values until a predetermined convergence criterion is met, indicating that the scores have stabilized.

6. **Assigning the Final PageRank Scores:** The algorithm then determines each webpage's final PageRank score based on importance after convergence is reached. The higher the PageRank score, the more important or significant a webpage is considered to be.

3.3.2 Text Rank Algorithm

We introduced the TextRank algorithm because it is widely used for summarizing text. Our objective is to create concise summaries from various domains of texts. TextRank helps us to figure out which words and sentences are the most important in a document by looking at how they are connected to each other. We begin with TextRank as it provides a solid foundation and serves as an excellent starting point. Once we comprehend TextRank, we will enhance our approach by considering the emotional tone of the text to develop our SentitextRank algorithm for even more effective summaries.

The TextRank algorithm is a graph-based ranking system designed to locate and extract critical information from text documents. TextRank rates the importance of each sentence in a document by examining the connections between words and sentences and determining each element's centrality in the graph.

Numerous natural language processing applications, including text summarization, keyword extraction, and document clustering, have found

widespread use for this approach. Moreover, it can help with text classification and information retrieval tasks, as well as identify topically comparable texts. Without the need for manual annotation or training, the TextRank algorithm can automatically create summaries of text documents, identify important keywords, cluster related documents, and extract important information from a vast collection of text data in the authors of [183].

In addition, TextRank is a versatile algorithm that can be easily integrated with other graph-based ranking systems, making it suitable for various text-processing tasks and applications in today's rapidly evolving information landscape. TextRank uses graphs to rank important elements of a text document based on their centrality in the work from authors [137].

In their 2004 publication "TextRank: Bringing Order into Texts", researchers Rada Mihalcea and Paul Tarau presented the TextRank algorithm, a novel tool for text summarization and keyword extraction. The late 1990s saw the introduction of PageRank, a graph-based ranking algorithm that was unsupervised by the authors of [118].

The TextRank method is a graph-based ranking algorithm designed to locate and extract critical information from text documents.

In summary, the TextRank method has shown to be a useful resource for natural language processing, facilitating document grouping, automatic summarization, and keyword extraction. It is graph-based technology, which draws inspiration from Google's PageRank approach, and makes it possible to extract significant information from texts without the requirement for training data or manual annotation.

Overall, TextRank is a graph-based ranking system that operates on multiple critical steps to extract significant information from text documents using the PageRank approach.

1. **Preprocessing**
   Divide the text into tokens (words or sentences). Processes may include removing stop words, stemming, and lemmatization.

2. **Graph Construction**
   Words or phrases are represented as nodes, and their relationships are shown by edges. Usually, the graph is weighted and undirected.

3. **Applying PageRank**
   Utilize PageRank methodology on the constructed graph. Calculate node importance iteratively using global information from the entire graph.

4. **Node Ranking**
   Rank nodes based on importance scores obtained from PageRank. Scores signify the significance of words or sentences in the document's context.

5. **Keyword Extraction**
   Identify top-ranked nodes as key elements or keywords in the document. These keywords represent crucial concepts for summarization and information retrieval.

TextRank algorithm efficiently leverages graph-based ranking mechanisms to extract crucial information from text documents, eliminating the need for manual annotation or training data.

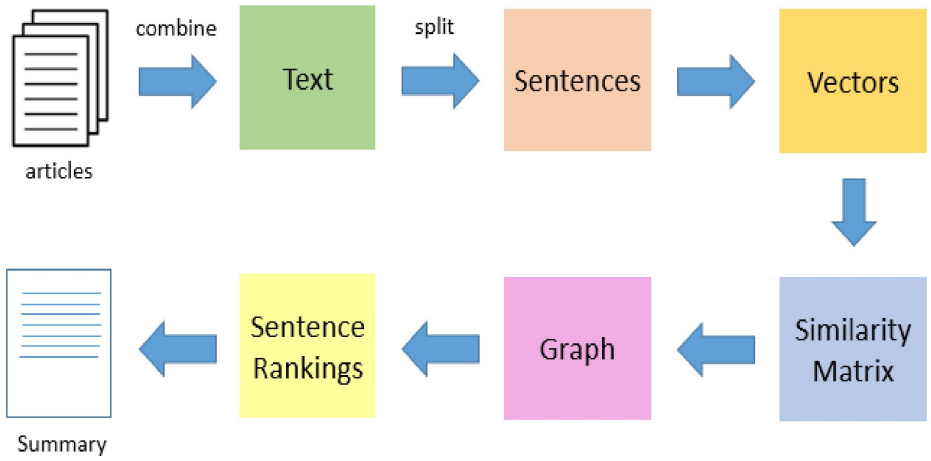


Figure 3.1: Architecture of TextRank Algorithm

3.3.3 BERT Sentence Embedding

When implementing the TextRank algorithm (see Figure 3.1), several crucial steps are involved in generating a meaningful summary. After combining and splitting the text into sentences, an important phase is to vectorize these sentences for measuring similarity. The choice of using BERT sentence embedding in this process is significant because BERT can capture intricate contextual relationships within sentences, providing a rich representation of the text. This enables more accurate measurement of sentence similarity

and enhances the effectiveness of the TextRank algorithm in extracting key information and generating coherent summaries.

Modern language models like BERT are often employed for a range of applications related to natural language processing. It is taught to anticipate masked words in a sentence using a deep bidirectional Transformer architecture.

BERT Sentence Embedding is a technique that utilizes the BERT language model to generate fixed-size vector representations (embeddings) of sentences. This allows sentences to be compared and analyzed based on their semantic similarity.

To produce BERT sentence embeddings, one usually takes the subsequent actions:

1. **Tokenization:** Tokenize the input sentences into individual words using WordPiece tokenization.

2. **Adding Special Tokens:** To indicate the beginning and finish of each sentence, place special tokens [CLS] at the beginning and [SEP] at the end.

3. **Conversion to Input Embeddings:** Convert the tokenized sentences into input embeddings using BERT's pre-trained weights. This involves mapping each word to its corresponding contextualized BERT embeddings, considering the surrounding words and their context.

4. **Deriving Sentence Embeddings:** There are various methods for acquiring the input embeddings and then deriving a fixed-size sentence embedding. Using the [CLS] token's embedding as the sentence representation, averaging the embeddings of every word in the sentence, or calculating the mean of the contextualized BERT embeddings for every token in the text are a couple of instances of these techniques in the work from authors [24].

3.3.4 Cosine Similarity

The TextRank algorithm involves key steps such as combining and segmenting text, followed by vectorizing sentences to facilitate similarity measurement.

The similarity matrix that serves as the foundation of the graph is measured using cosine similarity. This choice is strategic due to its effectiveness

in capturing the directional relationship between vectors, making it suitable for determining the similarity between sentence embeddings.

By leveraging cosine similarity, we ensure that the resulting graph accurately reflects semantic connections among sentences, ultimately enhancing the accuracy of TextRank in identifying and ranking key sentences for summarization.

A metric called cosine similarity is used to determine how similar two vectors or documents are to one another by the authors of [88].

The following formula can be used to determine cosine similarity:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|} \tag{3.1}$$

Equation 3.1 is the formula for cosine similarity.

The two vectors or documents being compared are denoted by $\mathbf{A}$ and $\mathbf{B}$. The dot product of $\mathbf{A}$ and $\mathbf{B}$ is indicated by $\mathbf{A} \cdot \mathbf{B}$. The magnitudes of vectors $||\mathbf{A}||$ and $||\mathbf{B}||$ represent the respective magnitudes of vectors $\mathbf{A}$ and $\mathbf{B}$, respectively in the work from authors [20]. It is frequently employed to determine how similar two vectors or documents are in the authors of [33].

3.4 Related Definition of used tools in Summarization

Pre-processing: Text data pre-processing entails a series of methods employed to ready text for analysis. The objective is to convert raw text into a more manageable and analyzable format by eliminating unnecessary elements, standardizing the text, and extracting important features. This process may encompass various actions, including: lowercasing all the text, removing punctuation and special characters, removing stop words, stemming or lemmatizing words, and normalizing unicode.

Lowercasing: Lowercasing is a crucial step in text pre-processing since it helps to standardize the text and gets rid of any potential inconsistencies that could occur from using different case for the same word than uppercase. It describes the procedure wherein all text is changed to lowercase letters. **Original sentence**: “I Love To Eat Pizza With My Friends ”.

After lowercasing: “i love to eat pizza with my friends”.

Stop word removal: Eliminating terms that are often used yet don’t add much to the text’s overall meaning. Stop words frequently occur in sentences like “and,” “the,” “for”. By removing unnecessary and noisy terms

from the text, stop words help to increase the effectiveness and precision of text analysis in the work from authors [58]. Removing common words that are unlikely to be useful for analysis.**Original sentence**: "I want to go to the store to buy some milk and bread." **After stop word removal**: "I want go store buy milk bread".

Gold summary: Gold summaries are manually written summaries that are regarded as superior reference summaries in the context of automatic text summarizing. The gold summary is typically created by human annotators who read and summarize the original text according to a set of guidelines or criteria.

Baseline summary: A baseline summary is a basic summary that is used to compare and assess how well more complex summarization systems perform in the context of automatic text summarization. Typically, a straightforward algorithm is used to generate the baseline summary. It chooses a portion of the original text's sentences or words based on parameters like frequency or position. Baseline summaries are frequently used to provide a baseline performance threshold that future systems must achieve or surpass. They also function as a benchmark for assessing the quality of other summarizing systems.

Compression ratio: The compression ratio compares the length of the summary to the source text to evaluate how concise a text summarizing method is. For instance, if the original text is 1,000 words long and the summary is 100 words long, then: In this case, you divide summary length or Original length to get a Compression ratio equal to 0.1 or 10%.

Order of sentence in the text: During tokenization and other preparation operations, the order of sentences is maintained. This allows for an accurate representation of the original text, ensuring that sequential information and context are maintained throughout the analysis process. The arrangement of sentences in a ordered form of a text is referred to as the summary's sentence order. The goal of abstractive summarization is to condense the most important information from the original text into a shorter amount of time. The quality of the summary can be greatly affected by the sequence of the sentences.

Content overlap similarity: Content overlap similarity measures the extent to which two segments of text share common elements, such as words or phrases. For instance, if the two sentences "The singing boy is wearing a blue hat" and "The boy with the blue hat is singing" are compared, there would be a high degree of content overlap similarity because, despite

differences in word order, they both share multiple words and communicate a comparable meaning. This closeness would help the TextRank algorithm create a strong link between the nodes in the graph that represent these sentences in the work from authors [118].

Unigrams:Unigrams refer to single words or tokens in a sequence of text. They are the simplest form of n-grams and represent individual units of text without considering their context. For example, in the sentence "NLP is the practice of understanding how people organize their thinking" the unigrams are " NLP,is, the, practice, of, understanding, how, people, organise, their, thinking" by the work from authors [114].

Bi-grams: Bigrams are sequences of two adjacent words or tokens in a sequence of text. They provide a bit more context than unigrams by considering pairs of words together.For example, in the sentence "NLP is the practice of understanding how people organize their thinking" the Bi-grams are "NlP is, is the, the practice, practice of, of understanding, understanding how, how people, people organise, organise their, their thinking" by the authors of [114].

Trigrams: Trigrams are sequences of three adjacent words or tokens in a sequence of text. They offer even more context than bigrams by considering sequences of three words together. For example, in the sentence "NLP is the practice of understanding how people organize their thinking," the trigrams are "NLP is the, is the practice, the practice of, practice of understanding, of understanding how, understanding how people, how people organise, people organise their, organise their thinking" in the work from authors [114].

N-grams: N-grams are sequences of n adjacent words or tokens in a sequence of text, where n can be any positive integer. They generalize the concept of unigrams, bigrams, and trigrams to consider sequences of any length. For example, if we consider n=4, then the 4-grams for the sentence "NLP is the practice of understanding how people organise their thinking" would include "NLP is the practice of understanding how people organise their thinking." N-grams are commonly used in natural language processing tasks such as text classification, language modeling, and machine translation by the work from authors [114].

3.5 Emotional Selection in the SentiTextRank Framework: Integrating Psychological Theory

The SentiTextRank framework carefully selected emotions based on a deep and comprehensive understanding of human emotional expression and their significant role in text summarization. The decision was not random; it stemmed from a thorough analysis considering various factors to ensure the effectiveness and relevance of the chosen emotions. Universality was a key consideration, with joy, sadness, anger, surprise, fear, and disgust being chosen for their universal recognition across diverse cultures and linguistic backgrounds. Furthermore, these emotions were deemed crucial for effectively encapsulating the essence of textual content in an inclusive manner. By incorporating these selected emotions into the SentiTextRank framework, it allows for a more nuanced and comprehensive understanding of the emotional content within the text.

Emotional diversity was also prioritized to capture the breadth and depth of human emotional experience. By encompassing a wide range of emotions, including positive ones like joy, negative ones such as sadness, anger, fear, and disgust, and neutral emotions like surprise in its repertoire, the algorithm seeks to mirror the intricate tapestry of emotions present in textual content.

The emotion hourglass model also guides the selection of emotions, suggesting that emotional experiences are dynamic and involve core affect cognitive appraisal, and expressive behavior. This highlights the temporal evolution of emotions influenced by individual differences and situational factors [31].

Each emotion was carefully assessed for its unique contribution to the overall emotional landscape of the text. For example, joy may signal moments of celebration or contentment, while sadness and anger might convey themes of loss or resentment. Surprise adds tension and anticipation, whereas fear evokes aversion or disgust. The theoretical underpinnings behind this selection can be found in the work of psychologist Paul Ekman in [56] on basic emotions which posits there exists core emotional states universally experienced by humans. Drawing upon this theoretical framework, the SentiTextRank leverages universality to enhance the summarization process.

In conclusion, the incorporation of carefully selected emotions in text summarization allows for a more nuanced and comprehensive understanding of the emotional content within the text.

3.6 Background Literature on Emotion Classification and Detection.

Technological approaches play a key role in unraveling human emotions. Researchers use advanced computational techniques to delve into emotion detection systems, analyzing facial expressions, speech signals, and textual cues to interpret emotional states. A selection of articles explores the latest advancements and challenges in this field.

The authors of [154] outline a four-phase approach for real-time human emotion detection using machine learning techniques. It leverages the MediaPipe face mesh algorithm and Principal Component Analysis to achieve an impressive 97% accuracy rate across diverse datasets. The framework is specifically tailored for robotic applications, providing an alternative method of understanding human emotions through facial expressions.

In the work of authors [23] Sentence-Level Emotion Detection Framework Using Rule-Based Classification the author introduces a rule-based system to detect emotions in text data like internet user reviews. The framework combines emotion theory with computational techniques, enhances sentiment analysis accuracy, and includes additional emotional cues such as emoticons and slang. It also involves creating an expanded emotion lexicon and algorithms capable of identifying formal and informal emotional signals. The study shows improved results compared to baseline methods and aims to provide resources for researchers focusing on sentiment analysis.

The authors of [174] extensively explore sentiment analysis and emotion identification. They highlight the importance of employing precise categorization models to avoid inaccurate results. They also investigate various emotion categorization models, proposes methods to enhance emotion research, and explores potential future advancements in the field.

The work from authors [76] introduce a new approach to identify emotions in text-based communications using a deep learning model based-on Long Short-Term Memory networks. The authors address challenges such as the absence of facial expressions and voice modulations, contextual understanding, sarcasm, and internet slang. Their proposed system leverages semantic and sentiment-based embeddings to detect various emotions in textual conversations. The method involves techniques for gathering datasets to train the model. The evaluation shows that the solution outperforms traditional Machine Learning baselines with potential practical applications including customer service monitoring.

The researchers of [63] explores emotion detection and analysis on social media, proposing a method to discern six emotions and their intensity using natural language processing and machine learning algorithms. The authors' model demonstrates significant accuracy in classifying tweets, achieving around 91.7% accuracy with J48 classifiers and 85.4% with SMO classifiers.

For detecting emotions from speech signals by extracting multiple acoustic features and using two classification models the authors of [151] introduces a method. Their methodology achieves an 80% accuracy for EmoDB and a 73% accuracy for RED, demonstrating its potential to improve human-machine interactions.

The work of authors from [65] explores advancements and challenges in sentiment analysis by introducing the novel aspect of emotion detection in Spanish texts. They provide an overview of the ninth edition of the Task on Semantic Analysis at SEPLN, held remotely on September 22, 2020, with fewer participants due to the COVID-19 pandemic. They highlight two significant subtasks: general polarity classification at three levels for identifying sentiment in Spanish tweets and a new challenge focusing on emotion detection from a newly compiled dataset. The authors summarize methodologies and findings from various participating teams, emphasizing linguistic nuances such as negation, irony, and sarcasm related to opinion subjectivity while also recognizing emotion analysis as an important area extending beyond opinion mining to encapsulate human emotional states reflected in Spanish social media content.

The authors of [166] offer an overview of the field of emotion recognition in speech. They discuss various features used for classifying emotions and reviews different models and techniques, including Deep Neural Networks, Support Vector Machines, Gaussian Mixture Models, and Convolutional Neural Networks. Through this review they also address challenges such as increased computational time with more neurons in DNNs and limited data size for unsupervised learning techniques.

In the dynamic landscape of emotion detection and classification, technological innovations intersect with real-world contexts. Emotion detection finds diverse applications across numerous domains, from enhancing customer experiences to addressing mental health challenges. This exploration delves into the practical significance and implications of emotion detection in various application domains, providing insights into how these technologies are shaping industries and enriching human interactions.

Exploring the landscape of emotion detection and classification through EEG brain-computer interface systems, the researcher of [8] offers a comprehensive evaluation of research encompassing 285 articles. These studies illuminate a wide spectrum of applications across communication, education, entertainment, and medicine The paper also discusses advancements in computational intelligence, machine learning, wireless EEG devices, and BCIs adapting to users' emotional states.

The work of authors from [78] discuss techniques for detecting emotions in microblogs and social media posts. They address the challenge of subjectivity and the ambiguous nature of emotions using a dimensional model to define emotion classes and propose a soft classification method. EmotexStream, a supervised learning system developed by the authors, is designed for real-time emotion tracking in text message streams and has applications in behavioral studies, public health, urban planning, as well as various emotion management applications. The challenges involve dealing with the casual style and semantic ambiguity of text messages along with the fuzzy boundaries of emotion classes.

The researchers of [3] delve into text-based emotion detection, emphasizing its significance, practical applications, current methodologies, and approaches, as well as the obstacles it encounters. They offer a deeper understanding of how emotions influence human interactions and explores the potential for businesses and individuals to leverage emotion detection for tailored services using customer emotions extracted from textual data such as reviews and ratings. Furthermore, they examine future research paths for emotion detection to enhance human-computer interactions and decision-making processes through the recognition of emotions in textual data.

The work of authors from [126] explore the fast-growing field of sentiment analysis and emotion detection in social media and text-based communication. They highlight the importance of processing unstructured data to understand human psychology, defining sentiment analysis as the process that assesses whether text expresses a positive, negative, or neutral viewpoint toward its subject. They also emphasize the significance of accurate sentiment analysis across various sectors such as business, stock market, healthcare, and education to improve products, services, and well-being. Additionally, they discuss how sentiment and emotion analysis has been essential in addressing challenges posed by events like the COVID-19 pandemic.

The authors of [32] provide a comprehensive overview of advancements in the field of affective computing, with an emphasis on emotion detection from text. They highlight the growing importance of emotion detection systems given the abundance of emotional data on the Social Web and outline potential applications such as suicide prevention and community well-being assessment. Various approaches in computational linguistics used to detect emotional states through text are explored, focusing on lexical and machine learning methods. The author categorizes significant works in emotion detection according to emotional models and employed approaches.

In the field of emotion detection and classification, researchers face challenges related to limited labeled datasets and processing emotional data. However, innovative solutions are emerging to address these obstacles and enhance emotion detection capabilities. Several articles explore these data challenges and highlight the developed approaches to overcome them, offering insights into efforts to decipher human emotions through data.

Addressing the challenge of limited labelled data for supervised learning models, the author elucidates the development of an automatic analysis system aimed at classifying sentiment and emotion in Twitter posts. The work of authors from [90] compiled a large-scale dataset comprising 17.5 million tweets, labelling them based on emojis that represent Ekman's six basic emotions. Traditional machine learning and advanced deep learning methods were applied to benchmark performance on this dataset, with a BiLSTM model delivering the best results: an F1-score of 70.92% for sentiment classification and 54.85% for emotion detection.

In the field of emotional detection and categorization, it is crucial to assess the performance, dependability, and efficiency of systems through evaluation and benchmarking. Researchers use thorough methods to compare various approaches, algorithms, and models to pinpoint strengths, weaknesses, as well as areas needing enhancement. This systematic assessment provides valuable insights into both the abilities and constraints of emotion detection systems that can direct future advancements in this area.

The Hourglass Model revisited by the authors of [160] explores emotion classification in AI and sentiment analysis. It introduces an updated version of the Hourglass of Emotions model for polarity detection and evaluates its effectiveness against three sentiment analysis datasets.

The researcher of [150] examines various methods developed for detecting human emotions through facial expressions, speech signals, physiological signals, and text semantics. They critically analyse the effectiveness of

different emotion detection approaches using multiple studies.

The work of researcher from [64] offers a comprehensive review of current technologies developed to detect human emotions. They explore various channels through which emotions can be read and emphasizes the importance of emotion detection in enhancing user experiences across different application domains. They also identify the strengths and limitations of existing technologies, while suggesting areas that need further research and development to advance the field of emotion detection.

In the rapidly advancing area of identifying and categorizing emotions, it is essential to take into account the ethical and social consequences. As these technologies are increasingly used in healthcare, marketing, and various other fields, concerns about privacy, consent, and prejudices emerge. Exploring these aspects of emotion detection allows for a better grasp of the wider impact on individuals and society. A variety of articles offer perspectives on the difficulties, debates, and factors related to these technologies - revealing the intricate relationship between technology and society in this domain.

The authors of [50] explore the evolution of linguistic analysis of suicide notes, from early manual approaches to modern natural language processing and machine learning techniques. They discuss the differentiation between genuine and fake suicide notes using a corpus collected by Shneidman in 1957, as well as recent advances focusing on automatic emotion analysis within these notes. The author also addresses the growing interest in the automatic detection of emotions on social media and its applications across various domains. They review methodologies including lexicon-based systems and supervised machine learning approaches like naive Bayes and support vector machines, highlighting how different aspects of language can contribute to sentiment expression in text. Overall, they present an overview of how advanced computational techniques can provide insights into suicidal behavior and mental health through emotion detection in text, specifically in suicide notes.

3.7 Overview of SentiTextRank

SentiTextRank, a variant of the TextRank algorithm, integrates sentiment analysis into the ranking process to prioritize sentences based on emotional categories such as joy, admiration, surprise, fear, disgust, anger, sadness,

and interest. The primary goal of SentiTextRank is to enhance applications like sentiment analysis, opinion mining, and emotion identification by providing valuable insights into the emotional intensity of text. By analyzing the sentiment of each sentence, SentiTextRank aims to offer a nuanced ranking that accurately reflects the emotional context, thereby improving the accuracy of sentiment analysis and opinion mining tasks and facilitating more informed decision-making across various domains in the work from authors [182].

Furthermore, SentiTextRank's integration of sentiment analysis contributes to a deeper understanding of textual data by prioritizing sentences based on their emotional content. This nuanced approach enables a more comprehensive interpretation of text, enhancing the overall analysis process and providing insights into underlying emotional categories. Its application in sentiment analysis of social media data, opinion mining in product reviews, and emotion detection in textual content has yielded promising outcomes, contributing to improved understanding and decision-making in diverse domains.

In conclusion, SentiTextRank aims to enrich the TextRank algorithm by incorporating sentiment analysis, offering a more comprehensive understanding of textual data while achieving the overarching goal of enhancing sentiment analysis and related tasks.

3.7.1 Sentiment lexicon based on the Emotion hourglass model used in SentiTextRank

The sentiment lexicon used by SentiTextRank is based on the Hourglass of Emotions model, which divides emotions into eight categories: joy, admiration, surprise, fear, disgust, anger, sadness, and interest. Each category represents a distinct aspect within the emotional spectrum and captures a wide range of emotional states found in text data.

This lexicon applies these emotional categories to assign sentiment scores to words and sentences according to their emotional associations. For example, words related to happiness and respect would receive higher sentiment scores indicating positive emotions, while words associated with fear, disgust, or anger would receive lower scores representing negative emotions.

SentiTextRank's sentiment analysis method enables it to prioritize sentences according to their emotional expression, offering valuable insights into the intensity of emotions conveyed in the text. Through scrutinizing each sentence's sentiment within the document's context, SentiTextRank

produces a refined ranking that effectively captures the emotional essence of the text.

Furthermore, the sentiment lexicon is continuously refined and updated to ensure its accuracy and effectiveness in capturing the nuances of human emotion. This iterative process involves incorporating insights from psychological research, linguistic analysis, and machine learning techniques to enhance the lexicon's coverage and granularity.

Overall, the sentiment lexicon based on the Hourglass of Emotions model serves as a foundational component of SentiTextRank, enabling it to achieve its overarching goal of enhancing sentiment analysis and related tasks by providing a comprehensive understanding of the emotional content within textual data.

3.7.2 SentiTextRank: Version 1

The SentiTextRank Version 1 presents a novel approach to summarizing text and categorizing emotions. This technique is useful for catching certain emotional nuances in written content and extracting important information in the field of natural language processing. This innovative technique combines sentiment analysis with TextRank algorithms for comprehensive emotion recognition and summarization of textual data. This version of SentiTextRank goes further than just summarizing by taking into account content and emotional context within the text. Version 1 seeks to offer summaries that encompass not only the information but also the underlying feelings expressed in the source material.

In Version 1 of SentiTextRank, the summarization process was primarily based on content similarity, utilizing techniques like TextRank to identify and rank sentences based on their structural and semantic relationships within the text.

The algorithm began by splitting the input text into individual sentences, classifying the sentences into different emotion categories, and then constructing a graph representation, where each sentence represented a node.

Next, edges were created for each group of emotion sentences based on their similarity scores, typically computed using cosine similarity or other metrics. The graph was then traversed for each group of emotion sentences using graph algorithms like PageRank to determine the importance of each sentence based on its connectivity to other sentences in the graph.

Finally, the top-ranked sentences were selected by prioritizing those with

the highest importance scores for each group of emotional sentences and merged them into a single summary considering the order of original sentences to form the summary.

In the following section 3.7.3, we provide a comprehensive breakdown of the Pseudo Code for SentiTextRank Version 1 along with relevant examples where applicable.

3.7.3 Pseudo Code for SentiTextRank Version 1

The Pseudo Code for SentiTextRank Version 1, offers a systematic approach to generating emotionally informed summaries from a source text. It categorizes sentences based on the Emotion Hourglass Model and crafts emotion-specific summaries using the TextRank algorithm. The final summary balances content with emotion, providing a concise representation of the source text.

The core principle of SentiTextRank Version 1 lies in its ability to transform a source text, denoted as ($Text_{Source}$), into a concise yet emotionally informed summary, labeled as ($Summary_{Final}$). This process is governed by a critical parameter termed the *compression ratio* (CR), which determines the proportional reduction from the source($Text_{Source}$) to the final summary ($Summary_{Final}$).

SentiTextRank Version 1 follows a strategic sequence of operations to achieve its objective. The source text is initially split into individual sentences, ensuring that each sentence represents a distinct unit of information.

Each of the split sentences is then categorized into distinct emotion categories (CAT_{em}) leveraging the SenticNet framework. These emotion categories encompass a wide range, including joy, admiration, surprise, fear, disgust, anger, sadness, and interest. Subsequently, for each emotion category (CAT_{em}), sentences classified under that category are gathered into a group ($SentenceSet_{em}$). These groups of sentences represent the emotional content of the source text.

Then, employing the TextRank algorithm, SentiTextRank Version 1 generates individual summaries ($Summary_{em}$) corresponding to each emotion category (CAT_{em}). These emotion-based summaries encapsulate the essence of the source(s) from varied emotional perspectives. Finally, the composition of the final summary ($Summary_{Final}$) amalgamates sentences proportionally chosen from each emotion-specific summary ($Summary_{em}$). This selection process, guided by the compression ratio CR, ensures the preservation of the original sequential arrangement present in the source

text ($Text_{Source}$) .

Algorithm 1: Pseudo Code for SentiTextRank Version 1

Function SentiTextRank Version 1($Text_{Source}$, CR):
- **Input:** $Text_{Source}$, CR (compression ratio)
- **Output:** $Summary_{Final}$
- Split $Text_{Source}$ into individual sentences
- Classify each sentence into an emotion category CAT_{em} based on the Emotion Hourglass Model, where $em \in$ {joy, admiration, surprise, fear, disgust, anger, sadness, interest}
- Gather all the sentences classified under the emotion category em into the set $SentenceSet_{em}$
- Apply the TextRank algorithm to $SentenceSet_{em}$ to create an emotion-specific summary $Summary_{em}$ for each emotion category em
- Initialize $Summary_{Final}$ as an empty list
- From each emotion-specific summary $Summary_{em}$, calculate the number of sentences to be selected using $num\_sentences_{em} = \text{round}(CR \cdot \text{length}(Summary_{em}))$
- Insert the selected sentences into the final summary $Summary_{Final}$ maintaining their order as they appear in $Text_{Source}$
- **return** $Summary_{Final}$

We will illustrate now the execution of the SentiTextRank version1 algorithm with a running example. We will compare the result with the gold summary reported below.

Gold summary:
Experts question if packed out planes are putting passengers at risk. U.S consumer advisory group says minimum space must be stipulated. Safety tests conducted on planes with more leg room than airlines offer.
The inputs for this example are $Text_{Source}$ (reported below, where total number of sentences in $Text_{Source}$ is 16 and length of Gold summary is 3) and the compression ratio $CR = 0.186$.

$Text_{Source}$**:**
Ever noticed how plane seats appear to be getting smaller and smaller? With

increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger? This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. "In a world where animals have more rights to space and food than humans," said Charlie Leocha, consumer representative on the committee. "It is time that the DOT and FAA take a stand for humane treatment of passengers." But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased . Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches. Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch. While most airlines stick to a pitch of 31 inches or above, some fall below this. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.

After Splitting source sentence we got the following sentences:

- *Ever noticed how plane seats appear to be getting smaller and smaller?*
- *With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk.*
- *They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger.*

- *More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger?*
- *This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans.*
- *"In a world where animals have more rights to space and food than humans," said Charlie Leocha, consumer representative on the committee.*
- *"It is time that the DOT and FAA take a stand for humane treatment of passengers."*
- *But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking?*
- *Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased.*
- *Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches.*
- *Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane.*
- *But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News.*
- *The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch.*
- *While most airlines stick to a pitch of 31 inches or above, some fall below this.*
- *While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches.*
- *British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.*

After classifying sentences we got followings sentences for each emotion:
$SentenceSet_{Joy}$: They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger? This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. "In a world where animals have more rights to space and food than humans," said Charlie Leocha, consumer representative on the committee." But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? While most airlines stick to a pitch of 31 inches or above, some fall below this. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches.

$SentenceSet_{Admiration}$: Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased . Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News.

$SentenceSet_{Surprise}$: "It is time that the DOT and FAA take a stand for humane treatment of passengers."

$SentenceSet_{Sadness}$: Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.

$SentenceSet_{Fear}$: no sentences

$SentenceSet_{Disgust}$: no sentences

$SentenceSet_{Interest}$: no sentences

$SentenceSet_{Anger}$: no sentences

After that, we obtain a summary for each category of emotions using TextRank taking into account the compression ratio CR for each category of emotions:

$Summary_{Joy}$: They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger.

$Summary_{Admiration}$: But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News.

$Summary_{Surprise}$: no sentences

$Summary_{Sadness}$: Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches.

$Summary_{Fear}$: no sentences

$Summary_{Disgust}$: no sentences

$Summary_{Interest}$: no sentences

$Summary_{Anger}$: no sentences

In this step, merging all emotion categories summaries $Summary_{em}$ and maintaining the order of the original text we got the final summary $Summary_{Final}$.

$Summary_{Final}$: They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News.

At last, we have measured RL-F1 score, Bert-sc, Cosim,Movsc, Pyrsc, Edig and Edio score for our generated summary compared with gold summary /reference summary. In case of this example we got RL-F1 score(0.247), Bert-sc (0.429), Cosim (0.478), Movsc(0.281), Pyrsc (0.139), Edig(2.519) and Edio (2.289).

3.7.4 SentiTextRank Version 2

SentiTextRank Version 2 is a noteworthy advancement in the field of sentiment-aware analysis and text summarization, offering a more sophisticated approach to understanding textual content.

This iteration of SentiTextRank goes beyond mere summarization by introducing a nuanced consideration of both content and emotional context within the text. By integrating these elements, Version 2 aims to provide summaries that capture not just the information but also the underlying sentiments expressed in the source material.

In Version 2, a significant enhancement was introduced by incorporating emotion similarity alongside content similarity in the summarization process. This involved several additional steps to capture and integrate emotional context into the summary generation.

First, an emotion analysis module was employed to extract emotional features from the input text, identifying the underlying emotional intensity such as joy, sadness, anger, or fear associated with each sentence. These emotional features were then used to compute emotion similarity scores between pairs of sentences, reflecting the degree of emotional similarity between them.

The algorithm then combined the content similarity and emotion similarity scores using a weighted approach, where the weights could be adjusted to emphasize either content or emotion more strongly based on the desired balance. This integration allowed the algorithm to generate summaries that not only preserved the factual content of the original text but also conveyed its emotional nuances.

Finally, the TextRank algorithm was applied to the combined similarity scores to select the most salient sentences for inclusion in the summary, ensuring that both content relevance and emotional resonance were considered in the final output.

Overall, the transition from Version 1 to Version 2 marked a significant evolution in SentiTextRank, enabling it to produce more comprehensive and emotionally resonant summaries by leveraging both content and emotion similarity in the summarization process.

In the subsequent section 3.7.5 we offer a detailed breakdown of the Pseudo Code for SentiTextRank Version 2, accompanied by relevant examples where applicable.

3.7.5 Pseudo Code for SentiTextRank Version 2

In Algorithm 2, we present a structured approach to generating summaries that merge content and emotion considerations. The algorithm starts by dividing the source text into individual sentences and then defines weight combinations to balance content and emotion similarity.

As like as SentiTextRank Version 1, we use a parameter known as the *compression ratio* (CR). This parameter controls how much of the original text(s) is reduced to produce the final summary $Summary_{Final}$, enabling customized and customized summarizing results.

Furthermore, an innovative concept is introduced through the definition of weight combinations of α and β. The condition of α and β weight combinations is $\alpha + \beta = 1$, where α represents content similarity, while β signifies emotion similarity.

Utilizing content embeddings from BERT, it calculates content similarity metrics for various combinations of (α) and generates emotion similarity metrics for different combinations of (β). These metrics are then combined to derive combined similarity metrics ($Similarity(s_i, s_j)$), which are subsequently normalized.

Applying the TextRank algorithm to these normalized metrics yields sentence scores $Similarity(s_i, s_j)$ and select the top scored sentences on the basis of CR to build $Summary_{Final}$. Here, we maintain the sentence order between the source text $Text_{Source}$ and generated summary ($Summary_{Final}$). This ensures that selected sentences in the summary align with their corresponding positions in the $Text_{Source}$, preserving contextual continuity and logical progression of ideas.

In essence, SentiTextRank Version 2 uniquely combines content similarity and emotion similarity measures to create extractive summaries that emphasize both content and emotion similarity.

Algorithm 2: Pseudo Code for SentiTextRank Version 2

Function SentiTextRank Version 2($Text_{Source}$, CR, α, β)**:**
  Input: $Text_{Source}$, CR (compression ratio), α, β
  with $\alpha + \beta = 1$ (α is the weight for content similarity and β is the weight for emotion similarity)
  Output: $Summary_{Final}$
  Split $Text_{Source}$ into individual sentences
  forall *pairs of sentences* s_i, s_j *in* $Text_{Source}$ **do**
    Compute $Content\_Similarity(s_i, s_j)$ as the cosine similarity of the BERT embeddings of s_i and s_j
    Compute $Emotional\_Similarity(s_i, s_j)$ as the cosine similarity of the vector of emotions as obtained by the Emotion Hourglass Model
    Set $Similarity(s_i, s_j) = \alpha \cdot Content\_Similarity(s_i, s_j) + \beta \cdot Emotional\_Similarity(s_i, s_j)$
  Normalize $Similarity(s_i, s_j)$ over i, j
  Apply the TextRank algorithm to compute the sentence score of each sentence, using $Similarity(s_i, s_j)$ and select the top scored sentences on the basis of CR to build $Summary_{Final}$, maintaining their order as they appear in $Text_{Source}$
  return $Summary_{Final}$

We will illustrate now the execution of the SentiTextRank version 2 algorithm with a running example. We will compare the result with the gold summary reported below.

Gold summary:
Experts question if packed out planes are putting passengers at risk. U.S consumer advisory group says minimum space must be stipulated. Safety tests conducted on planes with more leg room than airlines offer.

The inputs for this example are $Text_{Source}$ (reported below, where total number of sentences in $Text_{Source}$ is 16 and length of Gold summary is 3) and the compression ratio $CR = 0.186$.

$Text_{Source}$*:*
Ever noticed how plane seats appear to be getting smaller and smaller? With

increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger? This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. "In a world where animals have more rights to space and food than humans," said Charlie Leocha, consumer representative on the committee. "It is time that the DOT and FAA take a stand for humane treatment of passengers." But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking? Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased . Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches. Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch. While most airlines stick to a pitch of 31 inches or above, some fall below this. While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.

After Splitting source sentence we got the following sentences:

- *Ever noticed how plane seats appear to be getting smaller and smaller?*

- *With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk.*

- *They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger.*
- *More than squabbling over the arm rest, shrinking space on planes putting our health and safety in danger?*
- *This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans.*
- *"In a world where animals have more rights to space and food than humans," said Charlie Leocha, consumer representative on the committee.*
- *"It is time that the DOT and FAA take a stand for humane treatment of passengers."*
- *But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking?*
- *Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased.*
- *Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches.*
- *Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane.*
- *But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News.*
- *The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch.*
- *While most airlines stick to a pitch of 31 inches or above, some fall below this.*
- *While United Airlines has 30 inches of space, Gulf Air economy seats have between 29 and 32 inches, Air Asia offers 29 inches and Spirit Airlines offers just 28 inches.*

- *British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.*

After calculating content similarity metrics $Content\_Similarity(s_i, s_j)$ and emotion similarity metrics $Emotional\_Similarity(s_i, s_j)$, we combine these metrics for calculating $Similarity(s_i, s_j)$. Next, we apply the TextRank algorithm to obtain sentence scores for each sentence. Based on these sentence scores, we select the top sentences as our desired summaries. For example if we generate summary for content similarity (α) =0 and emotion similarity (β) =1. Then we get the following sentences as the generated summary $Summary_{Final}$ with the maintaining of original order.

$Summary_{Final}$ *(when alpha* $= 0$ *and beta* $= 1$*)*:
Ever noticed how plane seats appear to be getting smaller and smaller? Cynthia Corbertt, a human factors researcher with the Federal Aviation Administration, that it conducts tests on how quickly passengers can leave a plane. While most airlines stick to a pitch of 31 inches or above, some fall below this.

For this example our calculated values for Rouge-L-F1 (0.315), Bertscore (0.604), cosine similarity (0.318), mover score (0.279), pyramid score (0.249), and emotional distance with respective to gold summary Edio (2.570) and emotional distance with respective to original text Edig (2.474).

Similarly, we can generate summaries for different combinations of α and β, and then assess our generated summaries by calculating various evaluation metrics, including ROUGE-L-F1, BERTScore, cosine similarity, MoverScore, pyramid score, and emotional distance relative to both the gold summary and the original text. Subsequently, we compute the average for each evaluation metric individually.

3.7.6 Strengths and limitations of SentiTextRank

SentiTextRank is a variant of the TextRank algorithm specifically designed for sentiment analysis of texts.

Strengths:

1. Since it is an unsupervised approach, a labeled training dataset is not necessary.
2. It is easily integrated with different graph-based ranking mechanisms.

3. It considers both the positive and negative sentiments present in a text to determine its overall sentiment.

4. It incorporates linguistic information and rules, allowing it to effectively analyze short informal English texts.

Limitations:

1. SentiTextRank utilizes a pre-defined sentiment lexicon, like the hourglass of emotion models (e.g., joy, admiration, surprise, fear, disgust, anger, sadness, and interest). While this lexicon provides a foundation for sentiment analysis, it may not encompass all the diverse variations and nuances of sentiment, limiting the algorithm's ability to accurately capture the intricacies of emotional expression that extend beyond the predefined categories.

2. It may not accurately handle ambiguous or context-dependent sentiments, as it relies on a dictionary-based approach.

Overall, SentiTextRank demonstrates strengths such as being unsupervised, flexible in integration with different ranking mechanisms, considering overall sentiment, and incorporating linguistic information. However, it also has limitations such as reliance on a pre-defined lexicon, and challenges in handling ambiguous or context-dependent sentiments.

3.7.7 Evaluation metrics used in Summarization

When evaluating the effectiveness and quality of summarization techniques, evaluation measures are essential. The quality and accuracy of summarization techniques are evaluated using several measures. ROUGE, BERT Score, Cosine similarity, Mover score, and Pyramid score are a few evaluation metrics that are frequently used in text summarization. They are defined as follows:

Rouge: Rouge is a commonly used indicator to assess the quality of text summaries. Based on n-grams or the longest common subsequence, it calculates the overlap between the generated summary and the reference summary. This metric takes into account ROUGE-L, which measures the longest common subsequence, ROUGE-N, which reflects the length of an n-gram, and ROUGE-S, which measures skip-bigram similarity in the work from authors [187].

BERT Score: Through a comparison with reference summaries, the BERT Score is a metric used to assess the quality of generated summaries.

The generated summary and the reference summary's embedding similarity are determined using the BERT language model, which takes the summaries' content and fluency into consideration in the authors of [37, 51].

Mover Score: A tool for evaluation called the Mover Score compares the semantic similarity of two texts. It provides a thorough assessment of the quality of summaries by taking into account word embeddings and semantic meaning when calculating the similarity between generated summaries and reference summaries in the work from authors [94].

Pyramid Score: An assessment metric called the Pyramid Score evaluates the coherence and informativeness of summaries. It evaluates the quality of the generated summary by taking into account its structure and content and comparing it to a number of reference summaries by the authors of [132].

3.8 Dataset and experimental setup

To carry out an effective investigation into the SentiTextRank algorithm versions, it is essential to carefully select diverse sets of data representing different linguistic styles, content domains, as well as varying complexity levels to achieve reliable results from these experiments. Here we describe the datasets employed along with detailing our experimental setup devised for evaluating both version 1 and version 2 of SentiTextRank.

Datasets Utilized for SentiTextRank Version 1:

DUC 2001 & DUC 2004: NIST created datasets with 60 reference sets, 30 of which were designated for testing and 30 for training. Records, per-document summaries, and multi-document summaries were all included in each set, which was organized according to various standards including opinion and event sets.

CNN-DailyMail News Text Summarization: Over 300,000 distinct news stories from CNN and the Daily Mail are included in this dataset, which was first created for machine reading, comprehension, and abstractive question answering.

Datasets Employed for SentiTextRank Version 2:

WikiHow Summarization: Extracted from the WikiHow dataset comprising around 1000 documents, this dataset enables in-depth exploration specifically tailored for single-document summarization tasks.

BBC Articles: A dataset in .CSV format, sourced from BBC Articles, utilized in its entirety for single-document summarization tasks.

Blog Summarization: Human and algorithm-generated summarized data, although lacking detailed descriptions, was employed in its entirety for single-document summarization.

Tweets: Originating from Twitter, this dataset provides tweet summaries from various hashtags, encompassing diverse content for single-document summarization tasks.

Podcasts: This dataset, comprising over 100,000 transcribed podcast episodes, lacks ground truth summaries; however, episode descriptions serve as proxies for training supervised models in single-document summarization.

Hippocorpus: A dataset of 6854 English diary-like short stories, including recalled and imagined events, extensively utilized for single-document summarization tasks.

The experimental framework was designed using the full datasets for both SentiTextRank versions 1 (single and multi-document Summarization) and 2 (only single-document summarization), focusing solely on single-document summarization.

3.9 Experimental results for SentiTextRank: Version 1

The CNN/Daily Mail dataset (CNN) and the DUC2001 single document dataset (D01) are the two datasets used in the experiment for single-document summarization results in this section. The DUC2001 multi-document dataset (MD01) and the DUC2004 multi-document dataset (MD04) are the two datasets used for multi-document summarization results.

News datasets comprise the DUC 2001 single document and DUC 2001 multi-document datasets, which were gathered via the internet[1]. We used a sample data set of 54 documents for our single document studies.

The 50 items with multiple files and four reference files per item are included in the DUC 2004 multi-document summarization dataset[2], from which we used the first reference for each item. Furthermore, the CNN/Daily Mail dataset[3] was utilized.

[1] https://duc.nist.gov/data.html

[2] https://rb.gy/gp1gbt

[3] https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail

Three CSV files—test.csv, train.csv, and validation.csv—are included in this dataset. Three columns make up the test.csv file: id, article (which contains news article text in detail), and highlights (which is a summary of news article content). In this case, the "highlights" column serves as our experiment's reference summary. We ran our studies on the CNN/Daily Mail dataset's first 100 rows of text. Since the datasets only offer abstractive gold summaries, we changed the abstractive summaries into extractive summaries to enable a fair comparison. This procedure has been proposed in the work from authors [125, 84].

In contrast to the gold abstractive summary, an extractive reference summary ought to receive the greatest Rouge score. We take a greedy approach, adding one sentence at a time to the summary iteratively, making sure that the Rouge score of the current set of selected sentences is maximized in relation to the entire gold summary, because finding the globally optimal subset of sentences that maximize the Rouge score is computationally intractable. This approach is continued until the current summary set has no more potential sentences that could improve the Rouge score. The extracted reference summary for the assessment is then the subset of sentences that we currently have.

Table 3.1 reports an excerpt from the CNN dataset (Original Text) and the corresponding reference summary (Gold Abstractive Summary).

Moreover, Table 3.1 contains the corresponding generated reference extractive summary (Reference Extractive Summary), the Lead baseline (Lead), The random baseline (RB) generated text randomly, the text generated with the Original TextRank baseline (OTR), the SentiTextRank generated summary (STR1Cos) with cosine similarity and finally the SentiTextRank generated summary (STR1S) with Content Overlap similarity.

The experimental results of several summarization techniques based on SentiTextRank version 1, including the Lead baseline (Lead), The random base line (RB) generated text randomly, the text generated with the Original TextRank (OTR) and, finally, the SentiTextRank generated summary (STR1Cos) with cosine similarity and the SentiTextRank generated summary (STR1S) with Content Overlap similarity are shown in Table 3.2. These techniques were tested on a variety of datasets, including single-document datasets such as CNN and DUC-2001 (D01), as well as multi-document datasets like DUC-2001 (MD01) and DUC-2004 (MD04). Based on the compression ratio, we first chose the most important sentences from the original documents as a baseline.

Original Text	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk...
Gold Abstractive Summary	Experts question if packed out planes are putting passengers at risk. U.S consumer advisory group says minimum space must be stipulated. Safety tests conducted on planes with more leg room than airlines offer.
Reference Extractive Summary	Ever noticed how plane seats appear to be getting smaller and smaller? This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. The distance between two seats from one point on a seat to the same point on the seat behind it is known as the pitch.
Lead	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger.
RB	While most airlines stick to a pitch of 31 inches or above, some fall below this. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. But could crowding on planes lead to more serious issues than fighting for space in the overhead lockers, crashing elbows and seat back kicking?
OTR	In a world where animals have more rights to space and food than humans,"Said Charlie Leocha, consumer representative on the committee. Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased. But these tests are conducted using planes with 31 inches between each row of seats, a standard which on some airlines has decreased, reported the Detroit News.
STR1Cos	Between to that while inches and 28 from crowding not has seats, only overhead point reported pitch. Atlantic's inches, known our Detroit could safety Airways and The easyJet in seat tests health two conducted the on is treatment 31 decreased, take airlines Virgin pitch seat one - FAA the that for on a these elbows each "It lead between inches, crashing which has row of back planes DOT putting standard on is planes to inches, humane passengers"...
STR1S	Sergio Pellissier struck late on to boost Chievo 13 points clear of the relegation zone with a 1-0 victory at Cesena, while Atalanta opened a seven-point gap on the drop zone with a precious win. Authorities alleged Gardner tried to smuggle 30kg of methamphetamine out of China to Australia. Only, these bullets are traveling at hundreds of miles per hour.

Table 3.1: An excerpt of the Original Text from the CNN dataset.

Dataset	Algorithm	RL-F1↑	BERT-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
D01	LEAD	**0.564**	**0.719**	**0.615**	**0.472**	**0.481**	**2.211**	2.699
	RB	0.293	0.571	0.425	0.250	0.217	2.476	2.763
	OTR	0.312	0.584	0.481	0.267	0.257	2.379	**2.416**
	STR1Cos	0.190	0.445	0.374	0.159	0.093	3.168	3.336
	STR1S	0.113	0.448	0.270	0.109	0.064	3.305	3.354
CNN	LEAD	**0.665**	**0.771**	**0.664**	**0.515**	**0.557**	**1.752**	2.461
	RB	0.250	0.531	0.303	0.203	0.160	2.371	2.614
	OTR	0.322	0.576	0.397	0.254	0.230	2.078	**2.364**
	STR1Cos	0.168	0.430	0.256	0.124	0.079	3.324	3.503
	STR1S	0.091	0.389	0.135	0.071	0.042	2.988	3.157
MD01	LEAD	**0.667**	**0.771**	**0.690**	**0.560**	**0.590**	**2.028**	2.640
	RB	0.109	0.458	0.214	0.085	0.054	2.739	2.928
	OTR	0.126	0.473	0.371	0.116	0.085	2.332	**2.515**
	STR1Cos	0.123	0.420	0.199	0.065	0.048	3.917	4.065
	STR1S	0.046	0.387	0.151	0.034	0.024	6.987	6.955
MD04	LEAD	**0.485**	**0.668**	**0.559**	**0.378**	**0.384**	**2.202**	2.600
	RB	0.164	0.496	0.328	0.152	0.089	2.516	2.685
	OTR	0.205	0.517	0.391	0.182	0.129	2.365	**2.346**
	STR1Cos	0.138	0.430	0.264	0.109	0.061	3.446	3.500
	STR1S	0.118	0.432	0.224	0.103	0.056	2.616	2.583

Table 3.2: The experimental results of SentiTextRank Version 1. Here, Lead baseline (Lead), The random base line (RB), Original TextRank (OTR), SentiTextRank generated summary with cosine similarity (STR1Cos) and the SentiTextRank generated summary with Content Overlap similarity (STR1S). Similarly, Bert score (BERT-sc), Rouge-L-F1 (RL-F1), Cosine similarity (Cosim), Mover score (Movsc), Pyramid score (Pyrsc), Emotional distance with respect to gold summary (Edig), and Emotional distance with respect to original text (Edio).

Several metrics were used to assess the summaries: BERT-sc and Rouge-L-F1(RL-F1), Cosine similarity(Cosim), Mover score (Movsc), Pyramid score(Pyrsc), Emotional distance with respect to gold summary (Edig), Emotional distance with respect to original text(Edio). Recall Oriented Understudy for Gisting Evaluation, or ROUGE, is a commonly used tool for evaluating summarization quality. For the longest sequence of n-grams, Rouge-L computes ROUGE in the work from authors [104].

One metric that is frequently used in text categorization tasks is the BERT score. It measures how comparable the generated summary and the reference summary are at the token level (as previously mentioned in section 3.7.7).

Cosine similarity (Cosim), Mover score (Movsc), Pyramid score (Pyrsc) are also defined in Section 3.7.7).The calculation procedure of Emotional distance with respect to gold summary (EdiG) and Emotional distance with

respect to original text (Edio) explained clearly in Section 3.10).

In Table 3.2 of the experimental results for SentiTextRank Version 1, the metrics RL-F1, Bert-sc, Cosim, Movsc, and Pyrsc are marked with upward arrows (↑) to indicate that higher values suggest better performance. Conversely, the metrics Edig and Edio are marked with downward arrows (↓) to indicate that lower values suggest better performance.

The outcomes consistently suggest that the Lead technique is superior than the other methods in producing high-quality summaries across the datasets. For D01, CNN, MD01, and MD04, Lead specifically receives the highest scores in RL-F1 and BERT-sc, Cosim, Movsc, Pyrsc and Edig. In certain assessment metrics such as for Edio, the OTR approaches perform well.

It should be noted that the Lead method's superior performance can be attributed to the fact that all of the experiments were carried out using news datasets; in fact, this result is consistent with findings from the literature, which shows that on news datasets, a baseline made up of the leading sentences typically performs better than extractive and abstractive models by the authors of [107].

Still, we think that the comparison between OTR and STR1Cos shows encouraging results.

The fact that including emotions into a summary does not negatively impact TextRank performance indicates that it is possible to create a summary that accurately captures both the sentiment and the information of the original texts. In order to validate this intuition with experiments, we must establish an *emotional distance* between the summaries and the source documents (For calculating emotional distance we have conducted some experiments in section 3.10). In the future, we intend to further explore this point by (1) Using LLM, we analyze the emotional content encoded in both texts in order to quantify the emotional distance between generated summaries and source texts. By comparing probabilities or scores from the LLM model that represent the existence and strength of different emotions in the text and (2) taking into account human evaluation, we can quantify this emotional dissimilarity.

To fully grasp the usefulness of our suggested SentiTextRank (STR1Cos and STR1S) technique, additional study is required to assess its performance on another domain dataset.

3.10 Experimental results for SentiTextRank: Version 2

In our investigation of SentiTextRank Version 2, we have gathered diverse sets of data to thoroughly assess the algorithm's effectiveness across various text complexities.

In exploring SentiTextRank Version 2, a change in datasets was driven by several strategic considerations. Firstly we conducted the experiments for CNN and DUC 2001(Similar data of CNN and DUC 2001 used in SentiTextRank version 1). Then diverse datasets like WikiHow, blogs, BBC articles, tweets, podcasts, and Hippocorpus were used to assess the algorithm's performance across various domains. Secondly, transitioning to a larger dataset allowed for a more comprehensive evaluation of SentiTextRank Version 2. Lastly, the availability of reference summaries in larger volumes influenced the choice of datasets for a more robust evaluation framework.

Fo the exploration of SentiTextRank Version 2, a shift in datasets from the ones used in SentiTextRank Version 1, such as CNN, DUC-2001 for single document and DUC 2001 and 2004 Multi-document dataset, was motivated by several strategic considerations. Firstly, the adoption of diverse datasets like CNN, DUC 2001, wikihow, blog, bbc, tweets, podcasts, and hippocorpus aimed to scrutinize the algorithm's performance across a variety of domains, allowing us to observe patterns of results in different contexts.

These sets include the CNN dataset contains more than 300,000 distinct news articles from CNN and the Daily Mail. It was developed to assist in machine comprehension and abstractive question answering.The DUC 2001 dataset, created by NIST, includes 60 reference sets for testing and training. Each set contains records, per-document summaries, and multi-document summaries following various standards such as opinion and event sets. WikiHow Summarization with approximately 1000 documents from the WikiHow dataset, serving as a fundamental resource for single-document summarization tasks. We have also made use of BBC Articles in CSV format, Blog summary data containing both human-generated and algorithmic summaries, Tweets from different hashtags, transcribed Podcast episodes collection, and Hippocorpus featuring 6854 English diary-like short stories encompassing remembered and imagined events. We are able to assess the algorithm's performance on various kinds of content owing to this careful selection.

Our assessment involves a comprehensive examination using a range of algorithms, each with unique approaches. These encompass Lead (top few, exactly same number of sentences exists in reference summary), RB (Random Baseline), OTR (Original Text Rank), STR1Cos (SentiTextRank Version 1 with cosine similarity), STR1S (SentiTextRank Version 1 with content overlap similarity explained in section 3.4), STR2Cos (SentiTextRank Version 2 with cosine similarity), STR2S (SentiTextRank Version 2 with content overlap similarity).

To evaluate these algorithms, we use a variety of metrics. These include RL-F1 (Rouge-L), Bert-sc (Bert score), Cosim (cosine similarity), Movsc (Mover score), Pyrsc (pyramid score), Emotional distance with respect to gold summary (Edig), Emotional distance with respect to the original text(Edio).

In order to calculate the emotional distance (Edig), each sentence in the generated summary and gold summary is examined to determine the emotion vectors for each one. These vectors are based on emotion labels such as joy, admiration, surprise, fear, disgust, anger, sadness, and interest. Sentence lengths are normalized across both summaries for fair comparisons by using MinMaxScaler. The emotional distance is then calculated considering the average Euclidean distance among the identified emotional vectors of each sentence for each gold or reference summary and generated summary.

Similarly for calculating the emotional distance between generated summary and the original text (Edio), Initially, each sentence in the generated summary and original text is analyzed and identified the emotion vectors for each sentences based on emotion labels like joy, admiration, surprise, fear, disgust, anger, sadness, and interest. Sentence lengths are normalized across the generated summary and original text for fair comparisons by using MinMaxScaler. The emotional distance is then calculated considering the average Euclidean distance among the identified emotional vectors for each sentence of the generated summary and original text.

For all tables (3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, and 3.10) of the experiments on SentiTextRank Version 2, the metrics RL-F1, Bert-sc, Cosim, Movsc, and Pyrsc are marked with uprising arrows (↑) to indicate that the highest values among all methods suggest the best method. Conversely, the metrics Edig and Edio are marked with declining arrows (↓) to indicate that the lowest values suggest the best method.

This comprehensive method aims to rigorously assess SentiTextRank

Version 2 across different datasets and metrics, offering a deep understanding of its summarization abilities across diverse types and complexities of text.

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
CNN	LEAD	**0.665**	**0.771**	**0.664**	**0.515**	**0.557**	**1.752**	2.461
	RB	0.250	0.531	0.303	0.203	0.160	2.371	2.614
	OTR	0.322	0.576	0.397	0.254	0.230	2.078	**2.364**
	STR1Cos	0.168	0.430	0.256	0.124	0.079	3.324	3.503
	STR1S	0.091	0.389	0.135	0.071	0.042	2.988	3.157
	STR2Cos(wc=0)	0.518	0.687	0.531	0.414	0.429	1.887	2.465
	STR2Cos(wc=0.25)	0.274	0.550	0.330	0.234	0.197	2.099	2.406
	STR2Cos(wc=0.50)	0.277	0.551	0.332	0.237	0.201	2.103	2.412
	STR2Cos(wc=0.75)	0.273	0.550	0.328	0.236	0.198	2.102	2.410
	STR2Cos(wc=1)	0.259	0.543	0.319	0.214	0.185	2.203	2.449
	STR2S(wc=0)	0.272	0.542	0.325	0.220	0.205	2.382	2.625
	STR2S(wc=0.25)	0.326	0.577	0.398	0.259	0.236	2.101	2.389
	STR2S(wc=0.50)	0.313	0.569	0.383	0.249	0.222	2.119	2.394
	STR2S(wc=0.75)	0.315	0.573	0.388	0.250	0.225	2.094	2.381
	STR2S(wc=1)	0.322	0.576	0.397	0.254	0.230	2.078	**2.364**

Table 3.3: Experimental results focusing on SentiTextRank Version 2 of CNN Data. Here, Lead baseline (Lead), The random base line (RB), Original TextRank (OTR), SentiTextRank generated summary with cosine similarity (STR1Cos) and the SentiTextRank generated summary with Content Overlap similarity (STR1S). similarly, Bert score (BERT-sc), Rouge-L-F1(RL-F1),Cosine similarity(Cosim), Mover score (Movsc), Pyramid score(Pyrsc), Emotional distance with respect to gold summary (Edig), and Emotional distance with respect to original text(Edio).

Evaluating the experimental results for SentiTextRank Version 2 on CNN data, as presented in Table 3.3, reveals distinct strengths among different algorithms across various metrics. The LEAD algorithm consistently stands out as a top performer, achieving the highest scores in Rouge-L F1 (0.665), Bert-score (0.771), Cosine Similarity (0.664), Moverscore (0.515), Pyramid Score (0.557), and the lowest Edig (1.752). This highlights LEAD's superior ability to generate summaries that are both content-rich and semantically aligned with the original text. Additionally, STR2Cos with a content weight of 0 (wc=0) demonstrates remarkably strong performance, securing the second-highest results across several metrics: RL-F1 (0.518), Bert-score (0.687), Cosine Similarity (0.531), Moverscore (0.414), Pyramid Score (0.429), and Edig (1.887). This emphasizes the effectiveness of STR2Cos (wc=0) in producing summaries that are closely aligned with the original text, both in content and semantics. On the other hand, STR1Cos shows the weakest performance in terms of emotional distance, with the highest Edig (3.324) and Edio (3.503) scores, indicating a larger emotional gap from the reference summaries. Overall, the evaluations suggest that LEAD

and STR2Cos (wc=0) are particularly effective methods, each excelling in different aspects of summary generation, with LEAD leading in overall performance and STR2Cos (wc=0) showing substantial strength as well.

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
DUC-2001	LEAD	**0.564**	**0.719**	**0.615**	**0.472**	**0.481**	**2.211**	2.699
	RB	0.293	0.571	0.425	0.250	0.217	2.476	2.763
	OTR	0.312	0.584	0.481	0.267	0.257	2.379	**2.416**
	STR1Cos	0.190	0.445	0.374	0.159	0.093	3.168	3.336
	STR1S	0.113	0.448	0.270	0.109	0.064	3.305	3.354
	STR2Cos(wc=0)	0.482	0.673	0.540	0.392	0.407	2.258	2.532
	STR2Cos(wc=0.25)	0.312	0.583	0.429	0.274	0.257	2.347	2.412
	STR2Cos(wc=0.50)	0.312	0.584	0.429	0.274	0.257	2.348	2.413
	STR2Cos(wc=0.75)	0.306	0.581	0.425	0.269	0.251	2.355	2.394
	STR2Cos(wc=1)	0.233	0.546	0.378	0.198	0.178	2.432	2.436
	STR2S(wc=0)	0.261	0.544	0.357	0.211	0.200	2.582	2.574
	STR2S(wc=0.25)	0.306	0.578	0.449	0.266	0.248	2.367	2.394
	STR2S(wc=0.50)	0.296	0.574	0.442	0.253	0.238	2.359	2.400
	STR2S(wc=0.75)	0.301	0.577	0.443	0.259	0.243	2.350	2.386
	STR2S(wc=1)	0.312	0.584	0.481	0.267	0.257	2.379	**2.416**

Table 3.4: Experimental results focusing on SentiTextRank Version 2 of DUC-2001 Data.

Evaluating the experimental results for SentiTextRank Version 2 on the DUC-2001 dataset, as presented in Table 3.4, reveals notable insights into the performance of various algorithms across different metrics. The LEAD algorithm emerges as the top performer, achieving the highest scores in Rouge-L F1 (0.564), Bert-score (0.719), Cosine Similarity (0.615), Moverscore (0.472), Pyramid Score (0.481), and the lowest Edig (2.211), demonstrating its effectiveness in generating summaries that are both content-rich and semantically aligned with the original text.

Additionally, STR2Cos with a content weight of 0 (wc=0) shows strong performance across multiple metrics: RL-F1 (0.482), Bert-score (0.673), Cosine Similarity (0.540), Moverscore (0.392), Pyramid Score (0.407), and Edig (2.258). This demonstrates the effectiveness of STR2Cos in producing summaries aligned with the original text in terms of content and semantic coherence.

In contrast, STR1Cos exhibits the weakest performance in emotional distance, with high Edig scores (3.168) indicating a significant emotional gap from reference summaries. Similarly, STR1S also displays poor performance, as reflected by its weaker alignment with the emotional tone of reference summaries through its Edig scores (3.305).

Overall, LEAD and STR2Cos are particularly effective methods for summary generation on different aspects. LEAD excels overall while STR2Cos demonstrated substantial strength – providing valuable insights into their

effectiveness for generating comprehensive and semantically coherent summaries for the DUC-2001 dataset.

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
Blog	LEAD	**0.4889**	**0.6612**	**0.5710**	**0.4141**	**0.3905**	2.4920	2.5669
	RB	0.3135	0.5764	0.4601	0.2750	0.2202	2.5320	2.5521
	OTR	0.3637	0.6050	0.5486	0.3155	0.2967	2.3481	**2.3603**
	STR1Cos	0.2591	0.4496	0.4433	0.1825	0.1132	3.2510	3.2968
	STR1S	0.1288	0.4174	0.2631	0.1383	0.0733	2.7028	2.7515
	STR2Cos(wc=0)	0.4466	0.6471	0.5525	0.3880	0.3653	2.4105	2.4439
	STR2Cos(wc=0.25)	0.3468	0.5875	0.4985	0.3023	0.2700	2.4307	2.4058
	STR2Cos(wc=0.50)	0.3464	0.5876	0.4988	0.3022	0.2700	2.4306	2.4056
	STR2Cos(wc=0.75)	0.3493	0.5883	0.4971	0.3029	0.2698	2.4076	2.3975
	STR2Cos(wc=1)	0.3709	0.6003	0.5116	0.3032	0.2938	2.4635	2.4963
	STR2S(wc=0)	0.3808	0.6068	0.5149	0.3217	0.3015	2.4877	2.5332
	STR2S(wc=0.25)	0.3866	0.6182	0.5482	0.3141	0.3126	2.3566	2.3848
	STR2S(wc=0.50)	0.3906	0.6205	0.5531	0.3196	0.3171	**2.3406**	2.3737
	STR2S(wc=0.75)	0.3809	0.6148	0.5474	0.3130	0.3080	2.3424	2.3678
	STR2S(wc=1)	0.3637	0.6050	0.5486	0.3155	0.2967	2.3481	**2.3603**

Table 3.5: Experimental results focusing on SentiTextRank Version 2 of Blog Data.

In Table 3.5 of the document, when discussing the experimental results related to SentiTextRank Version 2 for Blog Data. Evaluating the algorithms across various metrics reveals that different methods excel in different aspects. LEAD consistently stands out as a top performer across several metrics, demonstrating its strength in producing summaries with high Rouge-L scores and strong cosine similarity with the original text.

Furthermore, STR2S with a content weight of 0.50 demonstrates remarkably balanced performance and secures a strong position on multiple metrics. This highlights the significance of content weight in determining the algorithm's effectiveness in aligning summaries with the original text while maintaining semantic coherence.

On the other hand, STR1Cos consistently exhibits greater emotional distance for Edig (3.2510) and for Edio (3.2968) from both gold summary(reference summary) demonstrating weaker performance.

Overall evaluations indicate that LEAD, STR2S (at content weight 0.50, where $\alpha + \beta = 1$ (see in section 3.7.4) and STR2Cos stand out as outstanding methods presenting superior performance based on diverse evaluation criteria - providing valuable insights into their effectiveness in generating comprehensive summaries encompassing varying characteristics.

In Table 3.6, an analysis of the experimental outcomes for SentiTextRank Version 2 on Podcast Data reveals distinct patterns across different measures. LEAD consistently holds a top position in various evaluation criteria,

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
Podcast	LEAD	**0.6730**	**0.7659**	**0.6663**	**0.5977**	**0.5655**	**2.3338**	3.3658
	RB	0.2089	0.4847	0.2482	0.2064	0.1306	3.3621	3.2728
	OTR	0.2288	0.4982	0.3020	0.2386	0.1593	2.9579	**2.8067**
	STR1Cos	0.1663	0.4514	0.2267	0.1485	0.0705	3.8727	3.8437
	STR1S	0.1104	0.4051	0.1490	0.1256	0.0477	3.5198	3.4574
	STR2Cos(wc=0)	0.6490	0.7545	0.6529	0.5845	0.5687	2.3735	3.3508
	STR2Cos(wc=0.25)	0.2011	0.4899	0.2776	0.2294	0.1403	3.0042	2.8438
	STR2Cos(wc=0.50)	0.2018	0.4902	0.2774	0.2292	0.1409	3.0071	2.8477
	STR2Cos(wc=0.75)	0.2014	0.4897	0.2744	0.2273	0.1409	3.0142	2.8545
	STR2Cos(wc=1)	0.2163	0.4913	0.2503	0.2050	0.1581	3.1701	3.0561
	STR2S(wc=0)	0.2351	0.5021	0.2732	0.2272	0.1773	3.3575	3.3228
	STR2S(wc=0.25)	0.2279	0.4972	0.2975	0.2389	0.1570	2.9773	2.8249
	STR2S(wc=0.50)	0.2282	0.4976	0.2988	0.2395	0.1570	2.9691	2.8143
	STR2S(wc=0.75)	0.2277	0.4982	0.3011	0.2390	0.1571	2.9586	2.8100
	STR2S(wc=1)	0.2288	0.4982	0.3020	0.2386	0.1593	2.9579	**2.8067**

Table 3.6: Experimental results focusing on SentiTextRank Version 2 of Podcast Data.

showcasing its ability to produce summaries with high Rouge-L scores, superior Bert scores, and strong cosine similarity with the original text.

An interesting discovery is that STR2S, with a content weight of 0.50 secures well-balanced performance; attaining the second-highest rank across multiple metrics, surpassing STR2Cos at content weight 0 and closely following behind STR2S at content weight 0.25. This observation emphasizes the crucial role of content weight in aligning the summary with the original text while maintaining semantic coherence.

Conversely, STR1Cos consistently demonstrates greater emotional distance from both Edig (3.8727) gold summary and original text Edio (3.8437), revealing its inadequacy in preserving emotional context. When compared, the performances highlight their effectiveness in creating diverse summaries as exemplary approaches, serving as valuable illustrations.

In Table 3.7, when analyzing the experimental outcomes related to SentiTextRank Version 2 applied to Wikihow Data, distinctive patterns emerge across various metrics. LEAD consistently maintains a dominant position across multiple evaluation criteria, signifying its capability to produce summaries with notably high Rouge-L scores (0.9672), superior Bert scores (0.9775), and robust cosine similarity with the original text (0.9657).

Conversely, STR2S with a content weight of 0.50 showcases a notably balanced performance, securing the second-highest position across several metrics. This method outperforms STR2Cos with a content weight of 0 and is closely followed by STR2S at a content weight 0.25, highlighting the critical role of content weight in optimizing the algorithm's efficiency. Specifically, STR2S(wc=0.50) exhibits comparative values: Rouge-L (0.1752), Bert-sc

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
Wikihow	LEAD	**0.9672**	**0.9775**	**0.9657**	**0.9554**	**0.9572**	**0.2096**	2.7545
	RB	0.1518	0.4745	0.1584	0.1784	0.0986	2.7066	2.7722
	OTR	0.1678	0.4830	0.1848	0.2055	0.1170	2.4714	**2.4777**
	STR1Cos	0.1008	0.3781	0.1232	0.1198	0.0461	3.5666	3.5723
	STR1S	0.0665	0.3899	0.0740	0.0930	0.0287	3.0090	3.0723
	STR2Cos(wc=0)	0.7312	0.8369	0.7337	0.7363	0.7153	1.3638	2.6803
	STR2Cos(wc=0.25)	0.1599	0.4872	0.1747	0.2062	0.1278	2.4853	2.5279
	STR2Cos(wc=0.50)	0.1622	0.4883	0.1768	0.2083	0.1298	2.4756	2.5199
	STR2Cos(wc=0.75)	0.1664	0.4909	0.1812	0.2110	0.1334	2.4631	2.5162
	STR2Cos(wc=1)	0.1766	0.5051	0.1859	0.2061	0.1442	2.5273	2.5852
	STR2S(wc=0)	0.2089	0.5105	0.2157	0.2334	0.1835	2.6033	2.7601
	STR2S(wc=0.25)	0.1740	0.4880	0.1885	0.2111	0.1259	2.4616	2.4868
	STR2S(wc=0.50)	0.1752	0.4887	0.1910	0.2130	0.1264	2.4540	2.4818
	STR2S(wc=0.75)	0.1748	0.4885	0.1911	0.2135	0.1265	2.4506	2.4781
	STR2S(wc=1)	0.1678	0.4830	0.1848	0.2055	0.1170	2.4715	**2.4777**

Table 3.7: Experimental results focusing on SentiTextRank Version 2 of Wikihow Data.

(0.4887), Cosim (0.1910), Movsc (0.2130), and Pyrsc (0.1264). Similarly, STR2S (wc=0.25) displays competitive values across these metrics: Rouge-L (0.1740), Bert-sc (0.4880), Cosim (0.1885), Movsc (0.2111), and Pyrsc (0.1259).

It is interesting to note that STR1S constantly exhibits a greater emotional distance from the original text as well as the gold summary, indicating its limited ability to capture emotional nuances effectively. Overall, LEAD, STR2S at content weight 0.50, and STR2S at content weight 0.25 emerge as the standout methods, exhibiting superior performance across diverse evaluation criteria, thereby offering valuable insights into their effectiveness in generating comprehensive summaries.

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
BBC Article	LEAD	**0.8397**	**0.8970**	**0.8427**	**0.7834**	**0.8063**	**0.8132**	2.3982
	RB	0.2023	0.5117	0.2240	0.1879	0.1387	2.3182	2.4464
	OTR	0.3003	0.5728	0.3323	0.2800	0.2440	2.0501	**2.2881**
	STR1Cos	0.1206	0.4232	0.1666	0.1098	0.0559	3.2107	3.3175
	STR1S	0.0588	0.3704	0.0665	0.0552	0.0244	3.7196	3.8079
	STR2Cos(wc=0)	0.4724	0.6798	0.4875	0.4394	0.4272	1.7534	2.3574
	STR2Cos(wc=0.25)	0.2557	0.5506	0.2785	0.2445	0.2080	2.1186	2.3342
	STR2Cos(wc=0.50)	0.2545	0.5498	0.2772	0.2431	0.2065	2.1191	2.3334
	STR2Cos(wc=0.75)	0.2454	0.5456	0.2697	0.2353	0.1975	2.1350	2.3347
	STR2Cos(wc=1)	0.1756	0.5105	0.2065	0.1626	0.1262	2.2996	2.3641
	STR2S(wc=0)	0.3280	0.5892	0.3489	0.3090	0.2851	2.0705	2.4209
	STR2S(wc=0.25)	0.3043	0.5751	0.3354	0.2861	0.2492	2.0436	2.2917
	STR2S(wc=0.50)	0.3058	0.5764	0.3374	0.2873	0.2508	2.0320	2.2869
	STR2S(wc=0.75)	0.3030	0.5746	0.3352	0.2837	0.2476	2.0421	2.2878
	STR2S(wc=1)	0.3003	0.5728	0.3323	0.2800	0.2440	2.0501	**2.2881**

Table 3.8: Experimental results focusing on SentiTextRank Version 2 of BBC Article Data.

Examining Table 3.8, which highlights the experimental results for SentiTextRank Version 2 when used with BBC Article Data, reveals patterns in a variety of measures that highlight different algorithmic performances.

LEAD is consistently the best at producing summaries with substantial cosine similarity with the original text (0.8427), high Rouge-L scores (0.8397), and superior Bert scores (0.8970). It also consistently retains its superiority across several evaluation metrics. Notably, STR2S, which has a content weight of 0.50, does well overall and comes in second place on several criteria. This approach shows its versatility by outperforming STR2Cos with a content weight of 0, closely followed by STR2S at a content weight of 0.25. The values Rouge-L (0.3058), Bert-sc (0.5764), Cosim (0.3374), Movsc (0.2873), and Pyrsc (0.2508) are particularly competitive in STR2S (wc=0.50). Likewise, STR2S(wc=0.25) presents values for the following metrics: Pyrsc (0.2492), Cosim (0.3354), Movsc (0.2861), Rouge-L (0.3043), and Bert-sc (0.5751).

However, STR1S method consistently communicates a greater emotional distance from its original material Edio (3.8079) in addition to the gold summary Edig (3.7196), indicating its inadequacies in accurately expressing subtle emotions. In general, LEAD, STR2S at content weight 0.50, and STR2S at content weight 0.25 exhibit exceptional performance in a variety of assessment criteria, providing useful insights into how well they produce thorough summaries.

Dataset	**Algorithm**	**RL-F1↑**	**Bert-sc↑**	**Cosim↑**	**Movsc↑**	**Pyrsc↑**	**Edig↓**	**Edio↓**
Tweets	LEAD	**0.7580**	**0.8465**	**0.7568**	**0.6499**	**0.6777**	**1.7027**	2.8193
	RB	0.3683	0.6157	0.3827	0.3174	0.3066	2.5880	2.9520
	OTR	0.3910	0.6364	0.4202	0.3437	0.3321	2.3004	**2.5345**
	STR1Cos	0.2106	0.5440	0.3345	0.1762	0.1087	3.0265	3.3987
	STR1S	0.0300	0.3675	0.0337	0.0273	0.0120	5.7121	5.8979
	STR2Cos(wc=0)	0.6697	0.7944	0.6778	0.5844	0.6081	2.0166	2.8187
	STR2Cos(wc=0.25)	0.3803	0.6286	0.4027	0.3439	0.3302	2.4404	2.7318
	STR2Cos(wc=0.50)	0.3797	0.6286	0.4016	0.3409	0.3289	2.4440	2.7305
	STR2Cos(wc=0.75)	0.3749	0.6264	0.3961	0.3352	0.3243	2.4512	2.7274
	STR2Cos(wc=1)	0.3553	0.6204	0.3703	0.2984	0.3036	2.6156	2.7912
	STR2S(wc=0)	0.4464	0.6637	0.4562	0.3776	0.3962	2.4935	2.8497
	STR2S(wc=0.25)	0.4089	0.6467	0.4345	0.3602	0.3523	2.2956	2.5547
	STR2S(wc=0.50)	0.4045	0.6428	0.4308	0.3541	0.3461	2.3058	2.5615
	STR2S(wc=0.75)	0.4056	0.6431	0.4312	0.3554	0.3464	2.2978	2.5587
	STR2S(wc=1)	0.3910	0.6364	0.4202	0.3437	0.3321	2.3004	**2.5344**

Table 3.9: Experimental results focusing on SentiTextRank Version 2 of Tweets Data.

SentiTextRank Version 2 experimental findings on Tweets Data are shown in Table 3.9. Various algorithmic performances are observed for the range

of metrics evaluated. LEAD continues to be the best on several assessment metrics, demonstrating its ability to produce summaries with excellent Rouge-L scores (0.7580), exceptional Bert scores (0.8465), and strong cosine similarity with the source text (0.7568). It's interesting to see that STR2S, which has a content weight of 0.50, does well overall, coming in second place on several criteria. This technique significantly outperforms STR2Cos, which has a content weight of 0, and STR2S, which has a content weight of 0.25. The values Rouge-L (0.4045), Bert-sc (0.6428), Cosim (0.4308), Movsc (0.3541), and Pyrsc (0.3461) are particularly competitive in STR2S(wc=0.50). Comparably, values for the following metrics are shown by STR2S(wc=0.25): Pyrsc (0.3523), Movsc (0.3602), Cosim (0.4345), Rouge-L (0.4089), and Bert-sc (0.6467).

On the other hand, STR1S consistently displays a more notable emotional distance from the original source Edio(5.8979) document as well as the gold summary Edig(5.7121), suggesting that it is not always possible to accurately capture subtle feelings. All things considered, LEAD, STR2S at content weight 0.50, and STR2S at content weight 0.25 stand out as exceptional performers, exhibiting great abilities across a variety of assessment criteria and providing insightful information about how well they can provide thorough summaries.

Dataset	Algorithm	RL-F1↑	Bert-sc↑	Cosim↑	Movsc↑	Pyrsc↑	Edig↓	Edio↓
Hippocorpus	LEAD	**0.6543**	**0.7394**	**0.5774**	**0.5882**	**0.5356**	**2.2375**	2.8802
	RB	0.2587	0.5001	0.2251	0.2420	0.1666	2.9186	2.8686
	OTR	0.2846	0.5130	0.2670	0.2633	0.1943	2.7147	**2.6515**
	STR1Cos	0.1858	0.4089	0.1852	0.1509	0.0786	3.8391	3.8079
	STR1S	0.0861	0.3781	0.0681	0.0779	0.0368	4.7539	4.7465
	STR2Cos(wc=0)	0.5342	0.6698	0.4764	0.4897	0.4427	2.4042	2.8393
	STR2Cos(wc=0.25)	0.2504	0.4993	0.2268	0.2435	0.1721	2.7984	2.7273
	STR2Cos(wc=0.50)	0.2510	0.4996	0.2269	0.2437	0.1723	2.7992	2.7276
	STR2Cos(wc=0.75)	0.2508	0.4994	0.2266	0.2429	0.1720	2.8002	2.7277
	STR2Cos(wc=1)	0.2598	0.5047	0.2268	0.2361	0.1801	2.8612	2.7901
	STR2S(wc=0)	0.2961	0.5253	0.2578	0.2751	0.2197	2.8489	2.8630
	STR2S(wc=0.25)	0.2897	0.5166	0.2696	0.2710	0.2002	2.7126	2.6582
	STR2S(wc=0.50)	0.2854	0.5140	0.2664	0.2663	0.1959	2.7164	2.6556
	STR2S(wc=0.75)	0.2836	0.5131	0.2654	0.2640	0.1946	2.7173	2.6552
	STR2S(wc=1)	0.2846	0.5130	0.2670	0.2633	0.1943	2.7147	**2.6515**

Table 3.10: Experimental results focusing on SentiTextRank Version 2 of Hippocorpus Data.

Whenever Table 3.10 is analyzed, tendencies emerge concerning different evaluation measures after SentiTextRank Version 2 is applied to Hippocorpus Data.

LEAD maintains a leading position and exhibits excellent performance in several assessment areas, such as Rouge-L (0.6543), Bert-sc (0.7394), and

Cosim (0.5774). On the other hand, STR2S, which has a content weight of 0.50, performs in a well-balanced manner and achieves a good rank in multiple measures. With a content weight of 0, this approach outperforms STR2Cos, whereas STR2S comes in second place with a content weight of 0.25.

The competition scores for Rouge-L (0.2854), Bert-sc (0.5140), Cosim (0.2664), Movsc (0.2663), and Pyrsc (0.1959) are specifically displayed by STR2S(wc=0.50). In a similar vein, STR2S(wc=0.25) displays values for the following metrics: Pyrsc (0.2002), Cosim (0.2696), Movsc (0.2710), Rouge-L (0.2897), and Bert-sc (0.5166). On the other hand, STR1S frequently presents a significant emotional distance from the source text as well as the gold summary, indicating shortcomings in the ability to accurately capture modest emotions.

Overall, the top three performers are LEAD, STR2S at content weight 0.50, and STR2S at content weight 0.25. They show excellent performance over a wide range of evaluation criteria and offer insightful information about how well they can create thorough summaries.

3.11 Discussion on Experimental Results

The thorough assessment of various summarization approaches, tested on a variety of datasets such as DUC-2001, CNN, DUC-2004, BBC Articles, Blogs, Tweets, Podcasts and Hippocorpus has provided crucial insights about the variations and effectiveness of these algorithms.

The evaluation of different summarization methods across various datasets has given us valuable insights. A crucial factor to take into account is the method by which we compute sentence similarity in order to determine how to assign weights to combination similarity (i.e., emotional and content similarity). These similarity between sentences, provide an additional layer of analysis beyond the TextRank model.

The key point is that these measures are independent of TextRank and can be seamlessly incorporated into the algorithm. This integration enhances the algorithm's adaptability by offering a different perspective on how sentences are related. By exploring these combined similarity (different combinations of content similarity and emotional similarity) measures, we can better tune the summarization model to different dataset scenarios, and comparisons making it stronger and more adaptable.

Furthermore, we found that some variations produce significant improvements over the original algorithm - this could have valuable implications for our work.

A comparison of SentiTextRank Versions 1 and 2 has yielded valuable insights, revealing significant advancements in emotion-considered summarizing techniques. The study demonstrated notable differences and improvements between the two algorithm versions. SentiTextRank Version 1 focused on cosine similarity and content overlap, whereas SentiTextRank Version 2 is more updated than Version 1. It considers both content similarity and emotion similarity alongside content relevance, resulting in consistently improved performance across numerous evaluation criteria for different datasets (refer to Tables 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, and 3.10).

The evaluation of various summarization approaches across a range of datasets, encompassing DUC-2001, CNN, DUC-2004, BBC Articles, Blogs, Wikihow, Tweets, Podcasts, and Hippocorpus yielded valuable insights into the effectiveness of these algorithms. When we looked at different methods, it was clear that the Lead method showed comparatively better performance, especially in datasets related to news (Table 3.3, 3.4, 3.5). This shows that it consistently delivers high-quality summaries for various datasets because for this method we are taking always top sentences. Although the Lead method is strong, our main interest lies in exploring Sentitextrank Versions 1 and 2 and their advancements. In most cases, our Sentitextrank Versions 1 and 2 outperform other methods, unlike the Lead method, which shows inferior performance.

Analyzing SentiTextRank Version 2 in greater detail revealed that content weight is essential for maximizing summarization results (Table 3.5). STR2S, with a content weight of 0.50, stood out as an example of balancing, proficiently handling a range of assessment criteria, and highlighting the significance of content weight in preserving coherence and ensuring that summaries are consistent with the source text.

Nevertheless, techniques exhibiting greater emotional distance from original texts as well as gold summaries frequently demonstrated inadequacies in accurately representing complex emotions (Tables 3.5, 3.7). This highlighted how important emotional context is in determining the overall quality and perceived usefulness of the summaries generated.

With STR2Cos at content weight 0 (STR2Cos(wc=0)), an unexpected finding emerged that prioritized emotional similarity above content relevance (Tables 3.5, 3.8). Surprisingly, these highly expressive summaries

demonstrated competitive performance on a range of assessment parameters even with little content overlap.

Tables 3.5 and 3.8 demonstrate the strong impact of emotional resonance on the perceived effectiveness of summary, challenging the traditional belief that content faithfulness alone determines summary quality.

Overall, if we observe Table 3.2, which presents the results of summarization experiments based on SentiTextRank Version 1, and Table 3.3, which presents experimental results focusing on SentiTextRank Version 2 on CNN data, and Table 3.4, which presents experimental results focusing on SentiTextRank Version 2 on DUC-2001 data, we notice a commonality between the CNN and DUC-2001 datasets. In these tables, all evaluation metrics such as RL-F1, Bert-sc, Cosim, Movsc, Pyrsc, Edig, and Edio show better results for SentiTextRank Version 2(STR2Cos(wc=0),STR2Cos(wc=0.25), STR2Cos(wc=0.50),STR2Cos(wc=0.75),STR2Cos(wc=1)and STR2S(wc=0), STR2S(wc=0.25), STR2S(wc=0.50), STR2S(wc=0.75), STR2S(wc=1)) compared with SentiTextRank Version 1(STR1Cos, STR1S).

Similarly, from Tables 3.5, 3.6, 3.7, 3.8, 3.9, and 3.10 we see that all evaluation metrics, such as RL-F1, Bert-sc, Cosim, Movsc, Pyrsc, Edig, and Edio, show better results for SentiTextRank Version 2(STR2Cos(wc=0),STR2Cos(wc=0.25),STR2Cos(wc=0.50), STR2Cos(wc=0.75),STR2Cos(wc=1)and STR2S(wc=0),STR2S(wc=0.25), STR2S(wc=0.50),STR2S(wc=0.75), STR2S(wc=1)) compared with SentiTextRank Version 1(STR1Cos,STR1S).

Therefore, after comparing the results, we can confidently conclude that SentiTextRank Version 2 outperforms SentiTextRank Version 1 across all evaluation metrics.

Based on our findings, it is apparent that the Lead method generally provides less emotional distance when we measure it in perspective of the gold summary. However, when considering our proposed Sentitextrank method, it consistently demonstrates even lower emotional distances, indicating its superior ability to capture emotional nuances effectively.

Excluding the Lead method from consideration, our Sentitextrank method emerges as the good performer in terms of emotional distance reduction. Furthermore, when all methods are taken into account, Sentitextrank consistently ranks as the second-highest method in most cases, further highlighting its robust performance. These results reinforce the efficacy of our proposed approach and its significant contribution to enhancing the

emotional quality of summaries. Thus, our study not only underscores the importance of integrating emotions into summarization but also positions Sentitextrank as a leading method in achieving this objective.

Ultimately, these findings highlight the balance that must be struck in summary creation between emotional expressiveness and content relevancy. Comprehensive material representation may be compromised by emotionally charged summaries, even though they might excel in certain criteria. In the future, efforts should strive to achieve a better balance to provide summaries that accurately communicate both factual particulars and sophisticated emotional accents contained in the original content.

3.12 Conclusion and Future Directions

The research investigated different methods for summarizing various datasets. The findings consistently showed the superiority of the Lead technique, especially for news-related data. SentiTextRank Version 2 outperformed Version 1 by incorporating emotional context along with content relevance. It became evident that content weight played a crucial role in maintaining semantic coherence and matching summaries to source texts. The ongoing challenge of balancing content accuracy and emotional impact in summaries is highlighted by the fact that approaches emphasizing emotions often struggle with factual representation but offer enhanced emotional resonance.

Going forward, there are a few directions in the field of text summarization that should be investigated. Improving the way that content and emotion interact in summaries is one interesting topic. One such area of research could be the optimization of the content weight parameter to better balance the expression of emotions and factual information. It would also be advantageous to create thorough measurements to assess emotional distance and how it affects the quality of summaries.

Improved specialized knowledge and contextual nuances integrated into algorithms like SentiTextRank could lead to more precise and emotionally engaging summaries for a broader variety of texts. To further enhance the understanding of text summary methods, future studies could investigate the application of sophisticated language models and human assessments to validate the emotional resonance in summaries.

Furthermore, broadening the scope of summarizing method evaluation

outside news-related datasets would offer a thorough grasp of the approaches' effectiveness in other settings. Evaluating these techniques' performance and adaptability on technical, scientific, or literary content may provide important new perspectives on their effectiveness and generalizability.

To improve the quality and application of summarizing algorithms, future work should focus on finding the ideal balance between factual correctness and emotional expressiveness in summaries, This involves considering various text genres and specific contextual settings, aligning with our specific contributions in SentiTextRank Version 1 and Version 2.

3.13 Major and Personal Contribution in sentiment-preserving extractive summarization chapter

Conceptualization and Objective Setting:

Initiated the sentiment-preserving extractive summarization chapter by outlining its objectives, research questions, and methodologies. This involved defining the scope of the research and identifying key areas of focus.

Development of SentiTextRank Algorithm:

Took a lead role in developing the SentiTextRank algorithm, which included the creation of both version 1 and version 2. This process involved brainstorming innovative approaches, designing algorithmic frameworks, and implementing the algorithms using appropriate programming languages and tools.

Data Collection and Analysis:

Led the data collection efforts by identifying relevant datasets across various domains pertinent to extractive summarization. Conducted extensive data analysis to uncover meaningful insights, patterns, and trends within the collected datasets. This analysis served as a foundation for developing and refining the SentiTextRank algorithm.

Publication and Dissemination:

Authored a conference article titled "Exploring sentiments in summarization: SentiTextRank, an Emotional Variant of TextRank," which detailed the research findings and innovations achieved with SentiTextRank version 1. This publication served as a platform to disseminate the research outcomes to the academic community and beyond [122].

Evaluation and Future Directions:

Played a pivotal role in evaluating the performance of the generated summaries produced by SentiTextRank version 1 and version 2. This evaluation process involved the application of various metrics and methodologies to assess the effectiveness and efficacy of the algorithms. Additionally, provided insights and recommendations for future research directions in sentiment-preserving extractive summarization, outlining areas for further exploration and improvement.

Chapter 4

Conclusion

This research explores two important areas: Sentiment Analysis in Medical Web Pages, and Sentiment-Preserving Extractive Summarization. Furthermore, by addressing these important fields, it makes a substantial contribution to emotional content analysis, computational linguistics, and natural language processing.

4.1 Summary of contributions

Our research has made significant contributions to extractive summarization techniques for different domain as like as medical data, news, stories , scientific articles etc. In our study, we conducted sentiment analysis on medical web pages to identify emotional nuances, addressing the challenges of maintaining these nuances in concise summaries, and refining sentiment-preserving extractive summarization.

Moreover, the evolution of SentiTextRank demonstrates progress in integrating emotional resonance into summaries while maintaining factual accuracy. These findings underscore the complexity of preserving emotional nuances in condensed text representations and highlight ongoing efforts to refine summarization techniques for effective communication across diverse informational scenarios.

Sentiment Analysis in Medical Web Pages:

This study's investigation of emotional patterns within digital medical content marks a significant development in comprehending the complex interaction between emotions and musculoskeletal pathologies. By using innovative methods such as emotional fingerprints, it reveals a deeper level of emotional content integrated into web pages about health issues.

A particularly noteworthy finding is the recognition of disgust as a crucial emotion, serving as a discriminating factor differentiating among various musculoskeletal conditions. This discovery illuminates the psychosocial aspects of how individuals perceive and interact with medical information on the internet. Grasping these emotional differences provides an important context to analyze user behavior, potentially influencing their information seeking behaviors and healthcare-related decision-making processes.

Sentiment-Preserving Summarization:

The focus of the research on techniques for summarizing sentiment represents a significant advancement in extractive summarization methods. In particular, the exploration of SentiTextRank and its evolution, along with innovative content-weight balancing strategies, marks notable progress in this field.

SentiTextRank Version 1 laid the foundation by integrating emotional aspects into summarization processes to preserve sentiments while condensing texts. However, it showed limitations in achieving a balanced representation of emotions and factual precision.

The progression to SentiTextRank Version 2 demonstrates an improvement by refining the algorithm's ability to incorporate emotional resonance into summaries while maintaining factual accuracy. It signifies an ongoing effort to strike a delicate equilibrium between depth of emotion and objective content within concise summaries.

The investigation of content-weight balancing strategies within SentiTextRank exemplifies efforts aimed at striking that balance effectively—a complex challenge incorporating emotional nuances while respecting conciseness in constraints—an essential element for facilitating efficient information consumption.

This study's exploration emphasizes the intricate challenge involved in developing sentiment-preserving summaries—attempts to capture emotional complexities while condensing large amounts of information into manageable formats. It underscores continuous efforts towards refining summarization techniques ensuring they embody both emotive essence and factual accuracy vital for effective communication and comprehension.

4.2 Key findings and insights

In our study, we carefully examine important discoveries to explore the

intricate balance needed to retain sentiments in summaries. In this investigation, we also reveal crucial insights about the complicated aspects of creating summaries that preserve sentiments.

Sentiment Analysis in Medical Web Pages:

The identification of disgust as a key emotion that distinguishes between various musculoskeletal conditions is a major advancement. This finding emphasizes not only the psychosocial differences among different health issues but also has potential implications for how patients interpret medical information online.

Understanding the impact of disgust on shaping emotional responses to medical content provides important insights into user perceptions and reactions. It highlights the importance of tailoring information-sharing strategies based on emotional nuances, potentially influencing how individuals engage with and comprehend medical information related to specific health conditions.

Sentiment Preserving Extractive Summarization:

Summarization techniques, such as the SentiTextRank Version 1 and the evolution to SentiTextRank version 2, indicate promising directions for summarizing emotional content. These methods emphasize the need to maintain factual accuracy while capturing emotional resonance in condensed text representations. The development of summarization methods is evident in the simplicity and effectiveness of the Lead technique in news-related data, as well as in the improved emotional depth displayed by SentiTextRank Version 2. However, this exploration also emphasizes a recurring difficulty: striking a balance between emotional nuances and factual accuracy in summaries - highlighting the complexity involved in preserving textual emotions.

4.3 Recommendations for further research

Investigate research directions to explore emotional aspects of medical web content through an exploratory study for the development of the SentiTextRank:, and improve methods for capturing emotions in brief summaries to enhance conversational systems and grasp the emotional context in healthcare content.

Sentiment Analysis in Medical Web Pages:

Expanding the scope of research to encompass comparative studies across languages could yield valuable insights into cultural variations in emotional

expression within medical contexts. Comparative analyzes in diverse linguistic environments and cultural backgrounds can illuminate how emotional nuances manifest, offering a comprehensive understanding of the influence of culture on emotional expressions related to health information.

Studying patient-generated content on social networks provides an abundant source for real-life integration of emotions. To identify emotional responses, attitudes, and behaviors regarding health-related information, research in this field may involve collecting and analyzing user-generated content from diverse social media platforms. This exploration could enhance our understanding of how individuals express emotions and engage with medical content in real-world social settings.

Sentiment-Preserving Extractive Summarization

Further exploration of summarization methods could focus on fine-tuning content weight parameters to achieve a balanced blend of emotional expressiveness and factual accuracy across various types of texts and fields. Improving strategies for balancing content weight within summarization algorithms has the potential to produce nuanced and contextually relevant summaries.

In addition, exploring the adaptability and effectiveness of these summarization techniques across different genres of text, such as technical, scientific, or literary works, can offer comprehensive insights into their performance. A thorough examination of their suitability in diverse contexts may provide valuable perspectives on optimizing them for various informational scenarios.

Bibliography

[1] Issam Aattouchi, Saida Elmendili, and Fatna Elmendili. Sentiment analysis of health care. In *E3S Web of Conferences*, volume 319, page 01064. EDP Sciences, 2021.

[2] Omar Younis Abdulhammed and Pshtiwan Karim. Sentiment analysis using svm-based sso intelligence algorithm. 11 2022.

[3] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

[4] Basant Agarwal and Namita Mittal. Semantic feature clustering for sentiment analysis of english reviews. *IETE Journal of Research*, 60(6):414–422, 2014.

[5] Basant Agarwal, Namita Mittal, Basant Agarwal, and Namita Mittal. Machine learning approach for sentiment analysis. *Prominent feature extraction for sentiment analysis*, pages 21–45, 2016.

[6] Taufiq Mohamad Ahmad and Nur Atiqah Sia Abdullah. A case study on social media analytics for malaysia budget. *International Journal of Advanced Computer Science and Applications*, 12(10), 2021.

[7] Samar H Ahmed, Khalid Tawfik Wassif, and Emad Nabil. Clustering based sentiment analysis using randomized clustering cuckoo search algorithm. *International Journal of Computer Science and Network Security*, 20(7):159, 2020.

[8] Abeer Al-Nafjan, Manar Hosny, Yousef Al-Ohali, and Areej Al-Wabil. Review and classification of emotion recognition based on eeg brain-computer interface system research: a systematic review. *Applied Sciences*, 7(12):1239, 2017.

[9] Murtadha Talib AL-Sharuee, Fei Liu, and Mahardhika Pratama. An automatic contextual analysis and clustering classifiers ensemble approach to sentiment analysis. *arXiv preprint arXiv:1705.10130*, 2017.

[10] Murtadha Talib AL-Sharuee, Fei Liu, and Mahardhika Pratama. Sentiment analysis: an automatic contextual analysis and ensemble clustering approach and comparison. *Data & Knowledge Engineering*, 115:194–213, 2018.

[11] Kashif Ali, Hai Dong, Athman Bouguettaya, Abdelkarim Erradi, and Rachid Hadjidj. Sentiment analysis as a service: a social media based sentiment analysis framework. In *2017 IEEE international conference on web services (ICWS)*, pages 660–667. IEEE, 2017.

[12] Tanveer Ali, David Schramm, Marina Sokolova, and Diana Inkpen. Can i hear you? sentiment analysis on medical forums. In *Proceedings of the sixth international joint conference on natural language processing*, pages 667–673, 2013.

[13] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484, 2019.

[14] Eissa M Alshari, Azreen Azman, Shyamala Doraisamy, Norwati Mustapha, and Mustafa Alkeshr. Improvement of sentiment analysis based on clustering of word2vec features. In *2017 28th international workshop on database and expert systems applications (DEXA)*, pages 123–126. IEEE, 2017.

[15] Abeer Alzuhair and Mohammed Al-Dhelaan. An approach for combining multiple weighting schemes and ranking methods in graph-based multi-document summarization. *IEEE Access*, 7:120375–120386, 2019.

[16] Manuela Angioni and Franco Tuveri. A semantic approach to the extraction of feature terms. 01 2011.

[17] Saud Anjum. Sentiment analysis of mobile product reviews. 06 2019.

[18] Md Siam Ansary. A hybrid approach for automatic extractive summarization. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 11–15. IEEE, 2021.

[19] Luca Anselma, Mirko Di Lascio, Dario Mana, Alessandro Mazzei, and Manuela Sanguinetti. Content selection for explanation requests in customer-care domain. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 5–10, 2020.

[20] Muhammad Zaky Aonillah, Hasmawati Hasmawati, and Ade Romadhony. Question entailment on developing indonesian covid-19 question answering system. 09 2022.

[21] Ioannis Arapakis, Mounia Lalmas, B. Barla Cambazoğlu, Mari Carmen Marcos, and Joemon M. Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. 03 2014.

[22] Dimaz Cahya Ardhi and Dwi Puspita Sari. Sentiment analysis of youtube comments: Potential indonesian presidential election candidates. 11 2022.

[23] Muhammad Zubair Asghar, Aurangzeb Khan, Afsana Bibi, Fazal Masud Kundi, and Hussain Ahmad. Sentence-level emotion detection framework using rule-based classification. *Cognitive Computation*, 9:868–894, 2017.

[24] Youheng Bai, Yan Zhang, Kui Xiao, Yuanyuan Lou, and Kai Sun. A bert-based approach for extracting prerequisite relations among wikipedia concepts. 10 2021.

[25] S Rahamat Basha, J Keziya Rani, and JJCP Yadav. A novel summarization-based approach for feature reduction enhancing text classification accuracy. *Engineering, Technology & Applied Science Research*, 9(6):5001–5005, 2019.

[26] Ramesh Chandra Belwal, Sawan Rai, and Atul Gupta. A new graph-based extractive text summarization using keywords or topic modeling. *Journal of Ambient Intelligence and Humanized Computing*, 12(10):8975–8990, 2021.

[27] Abdulkadir Abubakar Bichi, Ruhaidah Samsudin, Rohayanti Hassan, Layla Rasheed Abdallah Hasan, and Abubakar Ado Rogo. Graph-based extractive text summarization method for hausa text. *Plos one*, 18(5):e0285376, 2023.

[28] Mohammad Bidoki, Mohammad R Moosavi, and Mostafa Fakhrahmad. A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities. *Information Processing & Management*, 57(6):102341, 2020.

[29] Or Biran and Kathleen R McKeown. Human-centric justification of machine learning predictions. In *IJCAI*, volume 2017, pages 1461–1467, 2017.

[30] Sandipan Biswas, Shivnath Ghosh, and Sandip Roy. A sentiment analysis on tweeter opinion of drug usage increase by textblob algorithm among various countries during pandemic. *Int. J. HIT. TRANSC: ECCN. Vol*, 6(2A):1–9, 2020.

[31] Erik Cambria, Marco Grassi, Amir Hussain, and Catherine Havasi. Sentic computing for social media marketing. *Multimedia tools and applications*, 59:557–577, 2012.

[32] Lea Canales and Patricio Martínez-Barco. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43, 2014.

[33] Litan Cao, Huabing Wei, Zhi Huang, and Minglei Shi. An operating status analysis system of reactor equipment based on voiceprint recognition technology. 07 2022.

[34] Jorge Carrillo-de Albornoz, Javier Rodriguez Vidal, and Laura Plaza. Feature engineering for sentiment analysis in e-health forums. *PloS one*, 13(11):e0207996, 2018.

[35] Amartya Chakraborty and Subrata K. Bose. Around the world in 60 days: an exploratory study of impact of covid-19 on online global news sentiment. 10 2020.

[36] Omar Chamorro-Atalaya, Dora Arce-Santillan, José Antonio Arévalo-Tuesta, Lilia Rodas-Camacho, Genaro Sandoval-Nizama, Rosa Valle-Chavez, and Yadit Rocca-Carvajal. Text mining and sentiment analysis of teacher performance satisfaction in the virtual learning environment. 10 2022.

[37] CS Richard Chan, Charuta Pethe, and Steven Skiena. Natural language processing versus rule-based text analysis: Comparing bert score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights*, 16:e00276, 2021.

[38] Yisong Chen and Qing Song. News text summarization method based on bart-textrank model. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 2005–2010. IEEE, 2021.

[39] Zixuan Cheng and Shunli Guo. Automatic text summarization for public health wechat official accounts platform base on improved textrank. *Journal of Environmental and Public Health*, 2022, 2022.

[40] Minjee Chung, Eunju Ko, Heerim Joung, and Sang Jin Kim. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, 117:587–595, 2020.

[41] Mario Coccia and Saeed Roshani. Dynamics of research topics in cloud computing technology: Insights from methodology of entity linking and burst detection. 07 2023.

[42] Antonio R Damasio. *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt, 1999.

[43] Antonio R Damasio, Daniel Tranel, and Hanna C Damasio. Somatic markers and the guidance of behavior: theory and preliminary testing. 1991.

[44] Morena Danieli and Elisabetta Gerbino. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*, volume 16, pages 34–39, 1995.

[45] Amitava Das, Sivaji Bandyaopadhyay, and Björn Gambäck. The 5w structure for sentiment summarization-visualization-tracking. 01 2012.

[46] Sajib Dasgupta and Vincent Ng. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709, 2009.

[47] Vera Demberg, Andi Winterboer, and Johanna D Moore. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539, 2011.

[48] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27, 2015.

[49] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810, 2021.

[50] Bart Desmet and Véronique Hoste. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

[51] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[52] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*, 2021.

[53] Mirko Francesco DI LASCIO, Manuela Sanguinetti, Luca Anselma, Mana Dario, Alessandro Mazzei, Viviana Patti, Simeoni Rossana, et al. Natural language generation in dialogue systems for customer care. In *Proc. of Seventh Italian Conference on Computational Linguistics*, pages 1–6. CEUR, 2020.

[54] Pasquale Dolce, Davide Marocco, Mauro Nelson Maldonato, and Raffaele Sperandeo. Toward a machine learning predictive-oriented approach to complement explanatory modeling. an application for evaluating psychopathological traits based on affective neurosciences and phenomenology. *Frontiers in psychology*, 11:446, 2020.

[55] Soumi Dutta, Vibhash Chandra, Kanav Mehra, Sujata Ghatak, Asit Kumar Das, and Saptarshi Ghosh. Summarizing microblogs during emergency events: A comparison of extractive summarization algorithms. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, pages 859–872. Springer, 2019.

[56] Paul Ekman. Are there basic emotions? 1992.

[57] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Exploring clustering for multi-document arabic summarisation. In *Asia Information Retrieval Symposium*, pages 550–561. Springer, 2011.

[58] Yaohou Fan and Chetan Arora. Stop words for processing software engineering documents: Do they matter? 01 2023.

[59] Kuan Fang, Qi Zhang, Zhuoran Zhuang, and Zi-Ke Zhang. Making recommendations better: The role of user online purchase intention identification. In *2016 International Conference on Software Networking (ICSN)*, pages 1–4. IEEE, 2016.

[60] Susan Feldman. Nlp meets the jabberwocky: Natural language processing in information retrieval. *ONLINE-WESTON THEN WILTON-*, 23:62–73, 1999.

[61] Asbjørn Følstad and Marita Skjuve. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*, pages 1–9, 2019.

[62] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*, pages 145–156. Springer, 2019.

[63] Bharat Gaind, Varun Syal, and Sneha Padgalwar. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*, 2019.

[64] Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. Emotion detection: a technology review. In *Proceedings of the XVIII international conference on human computer interaction*, pages 1–8, 2017.

[65] Manuel García-Vega, Manuel Carlos Díaz-Galiano, MA García-Cumbreras, Flor Miriam Plaza del Arco, Arturo Montejo-Raéz, Salud María Jiménez-Zafra, E Martínez Cámara, César Antonio Aguilar, Marco Antonio Sobrevilla Cabezudo, Luis Chiruzzo, et al. Overview of tass 2020: Introducing emotion detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain*, pages 163–170, 2020.

[66] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[67] Bahar Gezici, Necva Bölücü, Ayca Tarhan, and Burcu Can. Neural sentiment analysis of user reviews to predict user ratings. 09 2019.

[68] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 international conference on digital health*, pages 121–125, 2018.

[69] Salvatore Graziani and Maria Gabriella Xibilia. Innovative topologies and algorithms for neural networks. 07 2020.

[70] Kathleen M Griffiths, Thanh Tin Tang, David Hawking, and Helen Christensen. Automated assessment of the quality of depression websites. 12 2005.

[71] Hanane Grissette and El Habib Nfaoui. Deep associative learning approach for bio-medical sentiment analysis utilizing unsupervised representation from large-scale patients' narratives. *Personal and Ubiquitous Computing*, pages 1–15, 2021.

[72] Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–11, 2019.

[73] Vaibhav Gulati, Deepika Kumar, Daniela Elena Popescu, and Jude D Hemanth. Extractive article summarization using integrated textrank and bm25+ algorithm. *Electronics*, 12(2):372, 2023.

[74] Christian Gulden, Melanie Kirchner, Christina Schüttler, Marc Hinderer, Marvin Kampf, Hans-Ulrich Prokosch, and Dennis Toddenroth.

Extractive summarization of clinical trial descriptions. *International journal of medical informatics*, 129:114–121, 2019.

[75] Dani Gunawan, Siti Hazizah Harahap, and Romi Fadillah Rahmat. Multi-document summarization by using textrank and maximal marginal relevance for text in bahasa indonesia. In *2019 International conference on ICT for smart society (ICISS)*, volume 7, pages 1–5. IEEE, 2019.

[76] Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*, 2017.

[77] Kamber L. Hart, Roy H. Perlis, and Thomas H. McCoy. What do patients learn about psychotropic medications on the web? a natural language processing study. 01 2020.

[78] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7:35–51, 2019.

[79] Lu He, Tingjue Yin, Zhaoxian Hu, Yunan Chen, David A. Hanauer, and Kai Zheng. Developing a standardized protocol for computational sentiment analysis research using health-related social media data. 12 2020.

[80] Lu Hong. Internet public opinion hotspot detection and analysis based on kmeans and svm algorithm. 08 2010.

[81] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618, 2013.

[82] Ken Hutchison and Soundar Kumara. Big data analytics-sentiment analysis of twitter data using clustering techniques. In *IIE annual conference. Proceedings*, page 2495. Institute of Industrial and Systems Engineers (IISE), 2013.

[83] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022.

[84] Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pages 2101–2110, 2017.

[85] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. Fine-tuning textrank for legal document summarization: A bayesian optimization based approach. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 41–48, 2020.

[86] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*, pages 895–906, 2018.

[87] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[88] Wei Jiang and Bharath K. Samanthula. N-gram based secure similar document detection. 01 2011.

[89] Salud María Jiménez-Zafra, M Teresa Martín-Valdivia, M Dolores Molina-González, and L Alfonso Ureña-López. How do we talk about doctors and drugs? sentiment analysis in forums expressing opinions for medical domain. *Artificial intelligence in medicine*, 93:50–57, 2019.

[90] Muhamet Kastrati, Zenun Kastrati, Ali Shariq Imran, and Marenglen Biba. Leveraging distant supervision and deep learning for twitter sentiment and emotion classification. *Journal of Intelligent Information Systems*, pages 1–26, 2024.

[91] Zied Kechaou, Ali Wali, Mohamed Ammar, Hichem Karray, and Adel M. Alimi. A novel system for video news' sentiment analysis. 03 2013.

[92] David Khanaferov, Christopher Luc, and Taehyung Wang. Social network data mining using natural language processing and density based clustering. In *2014 ieee international conference on semantic computing*, pages 250–251. IEEE, 2014.

[93] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. A conversational agent in support of productivity and wellbeing at work. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 1–7. IEEE, 2019.

[94] Abhishek Kumar. *Extractive Text Summarization.* PhD thesis, 2023.

[95] R Satheesh Kumar, A Francis Saviour Devaraj, M Rajeswari, E Golden Julie, Y Harold Robinson, and Vimal Shanmuganathan. Exploration of sentiment analysis and legitimate artistry for opinion mining. *Multimedia Tools and Applications*, pages 1–16, 2021.

[96] Salima Lamsiyah, Abdelkader El Mahdaouy, Saïd El Alaoui Ouatik, and Bernard Espinasse. Unsupervised extractive multi-document summarization method based on transfer learning from bert multi-task fine-tuning. *Journal of Information Science*, 49(1):164–182, 2023.

[97] Jacopo Lanzone, Cristina Cenci, Mario Tombini, Lorenzo Ricci, Tommaso Tufo, Marta Piccioli, Alfonso Marrelli, Oriano Mecarelli, and Giovanni Assenza. Glimpsing the impact of covid19 lock-down on people with epilepsy: A text mining approach. 08 2020.

[98] Jingxuan Li, Lei Li, and Tao Li. Multi-document summarization via submodularity. *Applied Intelligence*, 37:420–430, 2012.

[99] Jiwei Li and Eduard Hovy. Sentiment analysis on the people's daily. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 467–476, 2014.

[100] Junqing Li, Xiaolong Chen, Wei Niu, and Hongshi Sang. An improved multi-objective imperialist competitive algorithm for surgical case scheduling problem with switching and preparation times. 04 2022.

[101] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv:1703.01008*, 2017.

[102] ZhengMin Li. Research on brand image evaluation method based on consumer sentiment analysis. 05 2022.

[103] Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697, 2021.

[104] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[105] Han Liu, Siyang Zhao, Xiaolin Zhang, Feng Zhang, Junjie Sun, Hong Yu, and Xianchao Zhang. A simple meta-learning paradigm for zero-shot intent classification with mixture attention mechanism. 07 2022.

[106] Jingfang Liu, Jun Kong, and Xin Zhang. Study on differences between patients with physiological and psychological diseases in online health communities: Topic analysis and sentiment analysis. *International Journal of Environmental Research and Public Health*, 17(5):1508, 2020.

[107] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[108] Duy Khang Ly, Kazunari Sugiyama, Ziheng Lin, and Min-Yen Kan. Product review summarization based on facet identification and sentence clustering. *arXiv preprint arXiv:1110.1428*, 2011.

[109] Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das, and Apurba Sarkar. Graph-based text summarization using modified textrank. In *Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018*, pages 137–146. Springer, 2019.

[110] Kishore Kumar Mamidala and Suresh Kumar Sanampudi. A novel framework for multi-document temporal summarization (mdts). *Emerging Science Journal*, 5(2):184–190, 2021.

[111] Vinod L Mane, Suja S Panicker, and Vidya B Patil. Summarization and sentiment analysis from user health posts. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–4. IEEE, 2015.

[112] Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. *arXiv preprint cmp-lg/9712004*, 1997.

[113] Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, 2018.

[114] Sven Martin, Jörg Liermann, and Hermann Ney. Algorithms for bigram and trigram word clustering. *Speech communication*, 24(1):19–37, 1998.

[115] Alessandro Mazzei, Cristina Battaglino, Cristina Bosco, et al. Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th International Natural Language Generation conference (INLG 2016)*, pages 184–192. Association for Computational Linguistic, 2016.

[116] Michael Frederick McTear, Zoraida Callejas, and David Griol. *The conversational interface*, volume 6. Springer, 2016.

[117] Mahdi Gholami Mehr et al. Using adaboost meta-learning algorithm for medical news multi-document summarization. *Intelligent Information Management*, 5(06):182, 2013.

[118] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[119] Anupam Mondal, Erik Cambria, Dipankar Das, Amir Hussain, and Sivaji Bandyopadhyay. Relation extraction of medical concepts using categorization and sentiment analysis. *Cognitive Computation*, 10:670–685, 2018.

[120] Seonghyeon Moon, Gitaek Lee, Seokho Chi, and Hyunchul Oh. Automated construction specification review with named entity recognition using natural language processing. 01 2021.

[121] Alejandro Moreno and Carlos A Iglesias. Understanding customers' transport services with topic clustering and sentiment analysis. *Applied Sciences*, 11(21):10169, 2021.

[122] Md Murad Hossain, Luca Anselma, Alessandro Mazzei, et al. Exploring sentiments in summarization: Sentitextrank, an emotional variant of textrank. In *CEUR WORKSHOP PROCEEDINGS*, volume 3596, pages 1–5. CEUR-WS, 2023.

[123] Jin-Cheon Na and Wai Yan Min Kyaing. Sentiment analysis of user-generated content on drug review websites. *Journal of Information Science Theory and Practice*, 3(1):6–23, 2015.

[124] P Nagamma, HR Pruthvi, KK Nisha, and NH Shwetha. An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In *International conference on computing, communication & automation*, pages 933–937. IEEE, 2015.

[125] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[126] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81, 2021.

[127] Ali Naserasadi, Hamid Khosravi, and Faramarz Sadeghi. Extractive multi-document summarization based on textual entailment and sentence compression via knapsack problem. *Natural Language Engineering*, 25(1):121–146, 2019.

[128] Sristy Sumana Nath and Banani Roy. Towards automatically generating release notes using extractive summarization technique. *arXiv preprint arXiv:2204.05345*, 2022.

[129] Korawit Orkphol and Wu Yang. Sentiment analysis on microblogging with k-means clustering and artificial bee colony. *International Journal of Computational Intelligence and Applications*, 18(03):1950017, 2019.

[130] Ashish R. Panchal. A survey of web mining and various web mining techniques. 09 2019.

[131] Ho-Min Park and Jae-Hoon Kim. Stepwise multi-task learning model for holder extraction in aspect-based sentiment analysis. 07 2022.

[132] Rebecca J Passonneau. Formal and functional assessment of the pyramid method for summary content evaluation. *Natural Language Engineering*, 16(2):107–131, 2010.

[133] Julie Polisena, Martina Andellini, Piergiorgio Salerno, Simone Borsci, Leandro Pecchia, and Ernesto Iadanza. Case studies on the use of sentiment analysis to assess the effectiveness and safety of health technologies: a scoping review. *IEEE Access*, 9:66043–66051, 2021.

[134] Anita M Preininger, Brett South, Jeff Heiland, Adam Buchold, Mya Baca, Suwei Wang, Rex Nipper, Nawshin Kutub, Bryan Bohanan, and Gretchen Purcell Jackson. Artificial intelligence-based conversational agent to support medication prescribing. *JAMIA open*, 3(2):225–232, 2020.

[135] Ratish Puduppully and Mirella Lapata. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527, 2021.

[136] Sayyed Ali Rafiei, Hamid Sheikhzadeh, and Mohammad Sabbaqi. A new reduced-interference source separation method based on a complementary combination of masking algorithm and mixing matrix estimation. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 44:1529–1547, 2020.

[137] Mohammad Masudur Rahman and Chanchal K. Roy. Textrank based search term identification for software change tasks. 03 2015.

[138] Muhammad Rizky Ramadhan, Sukmawati Nur Endah, and Aprinaldi Bin Jasa Mantau. Implementation of textrank algorithm in product review summarization. In *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–5. IEEE, 2020.

[139] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630, 2021.

[140] Ehud Reiter. Natural language generation challenges for explainable ai. *arXiv preprint arXiv:1911.08794*, 2019.

[141] Ana Reyes-Menendez, José Ramón Saura, and Cesar Alvarez-Alonso. Understanding# worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach. *International journal of environmental research and public health*, 15(11):2537, 2018.

[142] Sumbal Riaz, Mehvish Fatima, Muhammad Kamran, and M Wasif Nisar. Opinion mining on large scale data using sentiment analysis and k-means clustering. *Cluster Computing*, 22:7149–7164, 2019.

[143] Ramon Gouveia Rodrigues, Rafael Marques das Dores, Celso G Camilo-Junior, and Thierson Couto Rosa. Sentihealth-cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. *International journal of medical informatics*, 85(1):80–95, 2016.

[144] Rajendra Kumar Roul and Jajati Keshari Sahoo. Sentiment analysis and extractive summarization based recommendation system. In *Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018*, pages 473–487. Springer, 2020.

[145] N. Rusli, Amiza Amir, Nik Adilah Hanin Zahri, and Rais Ahmad. Snake species identification by using natural language processing. 03 2019.

[146] Matthew E Sachs, Antonio Damasio, and Assal Habibi. The pleasures of sad music: a systematic review. *Frontiers in human neuroscience*, 9:404, 2015.

[147] Spyridon Samothrakis and Maria Fasli. Emotional sentence annotation helps predict fiction genre. *PloS one*, 10(11):e0141922, 2015.

[148] Abraham C Sanders, Rachael C White, Lauren S Severson, Rufeng Ma, Richard McQueen, Haniel C Alcântara Paulo, Yucheng Zhang, John S Erickson, and Kristin P Bennett. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of covid-19 twitter discourse. *AMIA Summits on Translational Science Proceedings*, 2021:555, 2021.

[149] Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Scalerandi Marco, Mana Dario, Simeoni Rossana, et al. Annotating errors and emotions in human-chatbot interactions in italian. In *The 14th Linguistic Annotation Workshop*, pages 1–12. Association for Computational Linguistics, 2020.

[150] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.

[151] Nancy Semwal, Abhijeet Kumar, and Sakthivel Narayanan. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE international conference on identity, security and behavior analysis (ISBA)*, pages 1–6. IEEE, 2017.

[152] Wafa Shafqat and Yung-Cheol Byun. A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis. *Sustainability*, 12(1):320, 2019.

[153] Chanakya Sharma, Samuel Whittle, Pari D Haghighi, Frada Burstein, and Helen Keen. Sentiment analysis of social media posts on pharmacotherapy: A scoping review. *Pharmacology Research & Perspectives*, 8(5):e00640, 2020.

[154] Ali I Siam, Naglaa F Soliman, Abeer D Algarni, Fathi E Abd El-Samie, and Ahmed Sedik. Deploying machine learning techniques for human emotion detection. *Computational intelligence and neuroscience*, 2022, 2022.

[155] Rameshwer Singh and Rajeshwar Singh. Applications of sentiment analysis and machine learning techniques in disease outbreak prediction–a review. *Materials Today: Proceedings*, 81:1006–1011, 2023.

[156] Shahid Husain Singh. Singh pk, shahid husain m. *Methodological study of opinion mining and sentiment analysis techniques, International Journal on Soft Computing*, 5(1):11–21, 2014.

[157] Jo Armour Smith, Heidi Stabbert, Jennifer J Bagwell, Hsiang-Ling Teng, Vernie Wade, and Szu-Ping Lee. Do people with low back pain walk differently? a systematic review and meta-analysis. *Journal of Sport and Health Science*, 11(4):450–465, 2022.

[158] Giorgos Stoilos, Szymon Wartak, Damir Juric, Jonathan Moore, and Mohammad Khodadadi. An ontology-based interactive system for understanding user queries. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 330–345. Springer, 2019.

[159] Darius Andrei Suciu, Vlad Vasile Itu, Alexandru Cristian Cosma, Mihaela Dinsoreanu, and Rodica Potolea. Learning good opinions from just two words is not bad. 01 2014.

[160] Yosephine Susanto, Andrew G Livingstone, Bee Chin Ng, and Erik Cambria. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102, 2020.

[161] Shan Suthaharan. Machine learning models and algorithms for big data classification. *Integr. Ser. Inf. Syst*, 36:1–12, 2016.

[162] Harshvadan Talpada, Malka N Halgamuge, and Nguyen Tran Quoc Vinh. An analysis on use of deep learning and lexical-semantic based sentiment analysis method on twitter data to understand the demographic trend of telemedicine. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–9. IEEE, 2019.

[163] Justin E Tang, Varun Arvind, Calista Dominy, Christopher A White, Samuel K Cho, and Jun S Kim. How are patients reviewing spine surgeons online? a sentiment analysis of physician review website written comments. *Global Spine Journal*, 13(8):2107–2114, 2023.

[164] Dehao Tao, Yingzhu Xiong, Zhongliang Yang, Yongfeng Huang, Jin He, and Kevin Song. An unsupervised extractive summarization method based on multi-round computation. *arXiv preprint arXiv:2112.03203*, 2021.

[165] Andrés Torres-Rivera. Detecting new word meanings: A comparison of word embedding models in spanish. 01 2020.

[166] Anjali Tripathi, Upasana Singh, Garima Bansal, Rishabh Gupta, and Ashutosh Kumar Singh. A review on emotion detection and classification using speech. In *Proceedings of the international conference on innovative computing & communications (ICICC)*, 2020.

[167] Taner Uçkan and Ali Karcı. Extractive multi-document text summarization based on graph independent sets. *Egyptian Informatics Journal*, 21(3):145–157, 2020.

[168] Ravi Kumar Venkatesh. Legal documents clustering and summarization using hierarchical latent dirichlet allocation. *IAES International Journal of Artificial Intelligence*, 2(1), 2013.

[169] David Vilares, Haiyun Peng, Ranjan Satapathy, and Erik Cambria. Babelsenticnet: a commonsense reasoning framework for multilingual

sentiment analysis. In *2018 IEEE symposium series on computational intelligence (SSCI)*, pages 1292–1298. IEEE, 2018.

[170] Marilyn Walker, Candace Kamm, and Diane Litman. Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3-4):363–377, 2000.

[171] Dingding Wang and Tao Li. Document update summarization using incremental hierarchical clustering. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 279–288, 2010.

[172] Xun Wang, Masaaki Nishino, Tsutomu Hirao, Katsuhito Sudoh, and Masaaki Nagata. Exploring text links for coherent multi-document summarization. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, 2016.

[173] Yili Wang, KyungTae Kim, ByungJun Lee, and Hee Yong Youn. Word clustering based on pos feature for efficient twitter sentiment analysis. *Human-centric Computing and Information Sciences*, 8:1–25, 2018.

[174] Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*, 79:35553–35582, 2020.

[175] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[176] Chih-Ping Wei, Yeng Chen, Chin-Sheng Yang, and Christopher C. Yang. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. 04 2009.

[177] Albert Weichselbraun, Philipp Kuntschik, Vincenzo Francolino, Mirco Saner, Urs Dahinden, and Vinzenz Wyss. Adapting data-driven research to the fields of social sciences and the humanities. 02 2021.

[178] Jeff Wu. Recursively summarizing books with human feedback. 01 2021.

[179] Avaneesh Kumar Yadav, Ashish Kumar Maurya, Rama Shankar Yadav, et al. Extractive text summarization using recent approaches: A survey. *Ingénierie des Systèmes d'Information*, 26(1), 2021.

[180] Chin-Sheng Yang, Chih-Ping Wei, and Christopher C. Yang. Extracting customer knowledge from online consumer reviews. 08 2009.

[181] Fu-Chen Yang, Anthony JT Lee, and Sz-Chen Kuo. Mining health social media with sentiment analysis. *Journal of medical systems*, 40:1–8, 2016.

[182] Sen Yang, Leyang Cui, Jun Xie, and Yue Zhang. Making the best use of review summary for sentiment analysis. 01 2020.

[183] Keedong Yoo. Intelligent and pervasive archiving framework to enhance the usability of the zero-client-based cloud storage system. 01 2014.

[184] Hongyuan Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120, 2002.

[185] Qian Zhai, Harianto Rahardjo, Alfrendo Satyanaga, Yiyao Zhu, Guoliang Dai, and Xueliang Zhao. Estimation of wetting hydraulic conductivity function for unsaturated sandy soil. *Engineering Geology*, 285:106034, 2021.

[186] Lu Zhang, Jialie Shen, Jian Zhang, Jingsong Xu, Zhibin Li, Yazhou Yao, and Lejiang Yu. Multimodal marketing intent analysis for effective targeted advertising. 01 2022.

[187] Ming Zhang, Chengzhang Li, Meilin Wan, Xuejun Zhang, and Qingwei Zhao. Rouge-sem: Better evaluation of summarization using rouge combined with semantics. *Expert Systems with Applications*, 237:121364, 2024.

[188] Pei-ying Zhang and Cun-he Li. Automatic text summarization based on sentences clustering and extraction. In *2009 2nd IEEE international conference on computer science and information technology*, pages 167–170. IEEE, 2009.

[189] Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. *arXiv preprint arXiv:1706.08476*, 2017.

[190] Chiara Zucco, Barbara Calabrese, Giuseppe Agapito, Pietro H Guzzi, and Mario Cannataro. Sentiment analysis for mining texts and social networks data: Methods and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(1):e1333, 2020.

Appendix A

Side PhD work: Anticipating User Intentions in Dialogue Systems

A.1 Introduction

Dialogue systems (DSs), also known as chatbots or conversational agents, have become increasingly important in various applications such as customer care, virtual assistants, and task completion. Understanding user intentions is crucial in dialogue systems as it directly impacts the decisions and policies made by the system [105]. Dialogue systems can have a variety of shapes and purposes. Based on their conversational focus, Grudin and Jacques proposed the following taxonomy of DSs: Intelligent assistants strive to keep talks brief regardless of the topic, task-oriented agents are made to address specific problems through shorter interactions, and virtual companions participate in open-ended chats on any subject [72].

Task-oriented dialogue systems have become increasingly popular among companies for customer interaction. These systems offer numerous benefits to both the company itself and its customers, as well as provide advantages to human operators in customer care [40, 93, 139]. Dialogue systems, also known as DSs, are accessible around the clock and can retain the context of a conversation for extended periods. This enables users to address their issues even if they become momentarily distracted from the support session. Additionally, DSs enable the gathering of unrestricted natural language text directly from customers. This text can then be analyzed using computational linguistics methods [116].

Therefore, the authors would offer a highly valuable resource of information regarding the expectations, preferences, and actions of customers. Customers often approach the DSs with different attitudes and expectations [86, 61], which are often not fulfilled by technology. The discrepancy between customer attitudes and technological capabilities emphasizes how crucial it is for dialogue systems to correctly predict users' intentions. When an agent provides users with correct understanding of their requests, pertinent responses, and clear communication of its capabilities, users have a positive experience [62].

Many users have high expectations for dialogue systems and treat them as if they were human operators. They frequently include contextual information that an automated system might not find useful or relevant in their long explanations of their circumstances and problems. When the agent fails to match their expectations or provide the intended solution, these users may become angry or dissatisfied [86].

They might end the chat as a result of this. To put it briefly, studies show that users' expectations affect how they interact with chatbots and, ultimately, how satisfied they are with the experience. One may have expectations regarding the chatbot's ability to accurately understand the user's intent, deliver information promptly, and offer pertinent and acceptable responses. In particular, it is essential for customer service that users expect to be informed about matters that are pertinent to them (such as unusual circumstances involving services they have subscribed to).

Several strategies have been proposed to address users' expectations in dialogue systems. These tactics entail enhancing the chatbot's capacity to produce accurate and pertinent clarifying queries, similar to interactive search systems. Additionally, the focus is on providing tailored responses that anticipate the needs or intentions of user information [13].

It is crucial to make clear that the term "intention" in this context does not apply to the user's specific purpose conveyed in utterances within task-oriented conversation systems, nor does it refer to the user's intention to make a purchase. In terms of dialogue systems, user intentions pertain to the goals or objectives that users have when they interact with a chatbot [59].

Understanding these user intentions is crucial for dialogue systems to effectively meet users' needs and deliver satisfactory experiences. Researchers recently looked into end-to-end task-completion neural dialogue systems for movie ticket booking. They also investigated how errors in the natural

language understanding module can impact system performance [101].

It has been emphasized that users often do not possess sufficient linguistic knowledge to meet their expectations. In such cases, it is important for dialogue systems to utilize domain context knowledge in order to better predict user intentions and enhance interactions [53].

By utilizing domain context knowledge, dialogue systems can better predict user intentions and enhance interactions. Understanding and predicting user intentions is crucial for building successful dialogue systems, especially in customer care scenarios.

This study presents a solution for addressing situations where users seek clarification on unusual circumstances related to their subscribed services, an area that has not been extensively studied in previous research. Our study involved analyzing nearly 3000 customer conversations with a dialogue system (referred to as COM-DS) used by a telecommunications company. Specifically, we focused on conversations where customers sought explanations from the dialogue system. Our findings reveal that approximately 5% of all conversations in our corpus pertained to requests for explanations, with half of these discussions centering around unexpected charges or fees on customers' telephone accounts. This scenario poses a significant risk for the company as it could potentially lead to customer churn. Therefore, accurately understanding and addressing user intentions in customer care dialogue systems is crucial for providing satisfactory experiences, improving system performance, and reducing the risk of customer churn.

We created a new dialogue system called GEN-DS to efficiently handle ambiguous or grammatically incorrect requests for explanations regarding charges on customers' phone accounts that were not recognized. GEN-DS differs from the other systems previously discussed in that it is not exclusively dependent on grammatically correct and comprehensible language input. Rather, it analyzes the user's phone balance's transaction history and makes a distinction between regular and unusual transactions.

Additionally, the model developed can differentiate between transactions that significantly impact the user's phone account from those that have negligible economic effects. By incorporating domain context knowledge and analyzing user transaction history, GEN-DS can accurately identify and address user intentions related to unrecognized charges on their telephone accounts in a personalized and efficient manner.

When creating and building GEN-DS, a key aspect is to provide users with synthetic and valuable answers when they inquire about unrecognized

charges on their phone accounts. Instead of presenting a lengthy and less timely response that lists all recent transactions, GEN-DS aims to offer immediate responses that are concise yet informative.

We believe that the techniques we have developed can be applied to other situations where users encounter unusual states in a system, service, or commodity and seek clarification from a dialogue system. When customers identify anomalies in their point balance and ask for clarification, an airline may employ this kind of technology to manage its frequent flyer program. Similarly, it can also be useful in handling accumulated discounts at stores or any scenario requiring the management of balances related to customer-company transactions involving points, money, or other affected factors.

In general, the ability to anticipate user intentions in customer care dialogue systems is crucial to provide satisfactory experiences and improve system performance. By incorporating transaction history and domain context knowledge, GEN-DS is able to accurately identify and address user intentions regarding unrecognized charges on their phone accounts.

The rest of this research is organized as follows: With an emphasis on our Methodological Approach and Tools, Section A.2 provides an in-depth analysis of the principles underlying our research methodology. Within section A.3 we engage in a comprehensive Discussion on Experimental Results, providing a thorough examination and interpretation of our findings. Section A.4 navigates through the Identified Challenges encountered in Anticipating User Intentions, shedding light on the obstacles faced during the research process. Looking towards the future, Section A.5 outlines our envisaged Future Directions, providing a roadmap for potential advancements in the field. Section A.6 encapsulates our study with a Conclusion, summarizing key takeaways and contributions. In Section A.7 elucidates major contributions made in Anticipating User Intentions in Dialogue Systems, providing a consolidated overview of the significance of our research. Lastly, in Section A.8 My personal contributions for the Anticipating User Intentions in Dialogue Systems

A.2 Methodological Approach and Tools

The main findings of this chapter are presented in this section. We detail the process of building a discourse corpus on explanation requests in the customer service domain in Section A.2.1. We give a formalization of the evidence idea to the customer transaction domain in Section A.2.2. We

describe GEN-DS, a DS created especially to handle requests arising from corpus analysis and generate evidence-based replies, in Section A.2.3. We present an approach in Section A.2.4 for creating experimental scenarios based on actual conversations. Section A.2.5 describes an experiment with users aimed at assessing the quality impact of GEN-DS for the case of explanation requests in the customer care domain, based on the situations. The results of the experiment are shown in Section A.2.6.

A.2.1 Building a Corpus of Explanation Requests in Customer-Care Dialogues Domain

To establish a foundation for our working hypothesis, we initially gathered a corpus of Italian conversations [149]. The additional material includes the annotation for the corpus. Due to confidentiality reasons, the actual content can only be accessed by the authors for research purposes and cannot be publicly shared. To find recurrent linguistic patterns in customer requests for explanations, the primary goal of the corpus analysis was to identify essential aspects in these interactions, such as the average number of turns per discussion and the average turn duration per user/agent.

As indicated in Section A.1, the dataset consists of real conversations between users and the Digital Service of a telecom provider. A selection technique was used to pick a portion of the discussions from a bigger sample that was gathered during a 24-hour period. This sample comprised interactions in which clients specifically asked for answers; relevant discussions containing words like “why” and “how come” were identified by a straightforward string-matching technique. The resulting corpus has 1540 turns altogether, or 142 talks with one or more messages from one party every turn. Each interaction has roughly 11 turns on average, with client turns lasting about 9 tokens and agent turns lasting about 38 tokens on average. It is important to note that the typical conversation time that was in this corpus deviates from previous studies which have shown task-oriented dialogues tend to be shorter. To understand the characteristics of customer requests for explanations within the field of customer service, a corpus of Italian conversations was gathered [47].

The Italian corpus of conversations was collected to analyze the characteristics of customer requests for explanations in the customer care domain.

In Figure A.1, the DS turns more slowly while the customer responds more briefly. The original Italian dialogue excerpt is on the left, while the English translation is on the right.

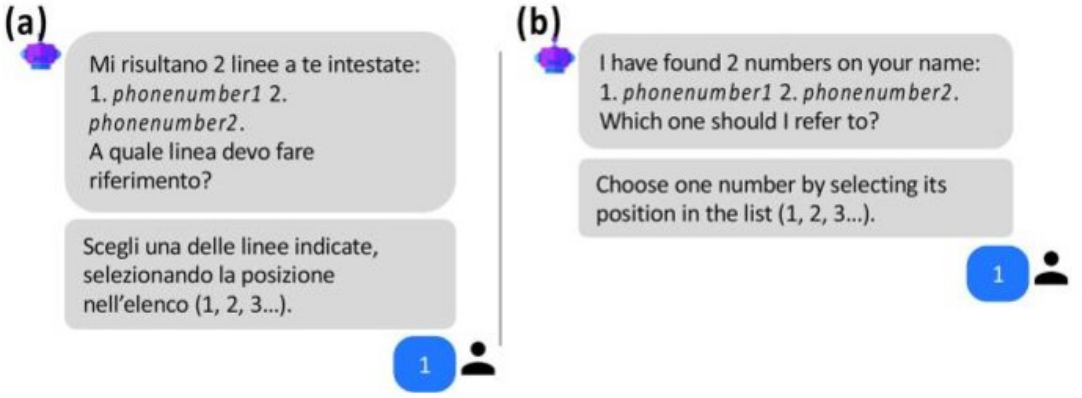


Figure A.1: An example of a DS-customer conversation.

This could be explained by the fact that a single turn typically comprises of multiple messages, particularly in the case of the DS. The average duration of client messages versus chatbot messages is another obvious difference. The current format of agent responses, which generally include comprehensive details like invoice items or possible options, can be blamed for this discrepancy. Customers' messages, however, are typically shorter. Customers' responses can sometimes only be simple yes/no responses or numerical digits (1, 2, etc.) that represent the alternatives the agent gave in the prior message (see Figure A.1).

These differences can lead to communication loops, in which requests or remarks are repeated several times within a single interaction between the chatbot and the user. Additionally, they might offer unrelated details that don't advance the conversational objective of giving the user a thorough explanation. Many commercial dialogue systems are designed with the assumption that the user's utterance contains relevant information [116].

Nevertheless, a preliminary examination of this corpus revealed that this assumption is sometimes untrue or only partially accurate. Linguistic input in user messages can be vague and difficult to follow at times (Examples 2 and 3), ambiguous at times (Example 1 - we provide a rough English translation to replicate the vagueness and ungrammatical nature of the original Italian utterances), or overly long and confusing at times. In these situations, the dialogue manager might have to overcome the absence of language specifics by asking for more clarifications or obtaining contextual data.

1. *Perché mi sono stati scalati dei soldi.*

(Why has some money been deducted from my account?)

2. *Salve. Vorrei sapere perché ho pagato 0,50 cent. Per sms se li ho gratis E i 2,00 euro in più per che cosa sono Grazie.*
   (Hello, I would like to know why I paid 0.50 cents for texts if I have them for free. And what about the additional 2.00 euros, what are they for? Thanks.)

3. *Bg come mai mi è addebitato altri euro ho qualche cosa attivato a pagamento.*
   (GM, how come I was charged other euros? Do I have something activated for a fee?)

4. *Come mai mi vengono addebitati costi di <serviceName> quando non è stato mai richiesto da me E come mai la bolletta è passata da 36 a 57 euro Ho già disdetto <serviceName> dai cellulari, mi sa che devo dare disdetta anche dal fisso poiché mi sento costantemente vessato e truffato dalla vostra compagnia. Inutile dire che è praticamente impossibile parlare con un operatore al telefono. Vergogna.*
   (Why am I being charged for <serviceName> when it has never been requested? And why has the bill increased from 36 to 57 euros? I have already canceled <serviceName> from mobile phones. I guess I will have to cancel it from the landline as well because I constantly feel harassed and cheated by your company. Needless to say that it is impossible to speak to an operator on the phone. Shame on you.)

5. *Scusami, ma vorrei sapere come mai mi vengono fatti certi addebiti.*
   (Sorry, but I would like to know why there are some charges.)

6. *Salve, vorrei sapere perché mi sono stati presi 12€ invece che dieci dall'ultima ricarica.*
   (Hi, I would like to know why you charged me 12€ instead of ten from the last top-up.)

7. *Buongiorno, vorrei sapere perché ho il credito in negativo, nonostante abbia fatto una ricarica da 15€ proprio stamattina.*
   (Good morning, I would like to know why I have a negative balance, despite I made a 15€ top-up just this morning)

Annotation experiments were conducted on the corpus prior to the creation and development of GEN-DS, in an effort to investigate potential recurrent patterns that might underlie the interactions between users and

chatbots (a more thorough explanation of the annotation scheme can be found in [149].

The annotation experiments conducted on the corpus helped identify three main categories for users' explanation requests to make GEN-DS design and development easier.

Category I: (58% of the occurrences in the corpus) a charge in the account is claimed, but no further information is provided (see Examples 1, 3, and 5).

Category II: (31% of the occurrences) the customer asks for an explanation about a charge providing vague information (Examples 2, 4, and 6).

Category III: (11% of the occurrences) the customer asks for an explanation about a negative balance (Example 7).

Thus, the corpus analysis demonstrated that, in customer service encounters, user requests may be imprecise and insufficiently descriptive, and they may be associated with (at least) one of the main categories we previously discussed. In considering this, we created a new DS that can generate a response to the request (see Section A.2.3) by utilizing typical symbolic NLG approaches and domain knowledge (see Section A.2.2). Based on the three categories found in the corpus, we assessed this DS (see Section A.2.4).

A.2.2 Importance, Effect, and Evidence in Relational Domain-Context Knowledge

The necessity of linking domain-specific data with accurate linguistic explanations has received a lot of attention nowadays [140].

Content selection is an important aspect of this task, as it determines the type of information that should be conveyed to the user. Symbolic, statistical, and neural methods are among the ways that have been proposed to achieve this goal (see [135] for a recent article on the state of the art).

In this paper, we adapt Biran and McKeown's proposed choice of content methodology [29]. Statistical classifiers are examined in Biran and McKeown's original proposal, particularly linear SVMs based on linear discriminant functions. They present the idea of "importance" which is the weight of a feature in categorizing every instance of the training set into the same class, and "effect" which is the weight of a feature in classifying a single data instance into a certain class. To determine and rank the elements that should be included in an NLG system for representing financial stock price patterns, the authors propose integrating these concepts into a single

concept termed "evidence". Biran and McKeown's work suggests that the notion of "evidence" plays a crucial role in identifying the characteristics that a natural language generation system should have to describe financial stock price patterns.

The authors suggest combining these two ideas into a single concept known as evidence. They demonstrate how the use of evidence can be utilized to prioritize and choose which features, in an NLG system, should be presented to users to describe the trends in financial stock prices. Specifically, important values are positions in narratives that "... represent semantically clear concepts that non-experts readily understand and are rooted in the true details of the prediction" [29].

Features that show significant importance and effect that are relevant in both the training set classifications and the current classification task are referred to be normal evidence. However, exceptional evidence refers to features that have a strong impact on the current classification task, but may not be as important in overall training set classifications.

Table A.1: Transaction Patterns and Impact Analysis by Category.

$DC-K-1$	M_1	M_2	M_3	M_4	M_5	M_6	M_7
S_1	9.99	9.99	9.99	9.99	9.99	9.99	9.99
S_2	0	0	0	0	2	2	2,2
S_3	0	0	0	0	0	0	1.59
$DC-K-2$	M_1	M_2	M_3	M_4	M_5	M_6	M_7
S_1	10	10	10	10	10	10	10
S_2	0	0	0	0	0	2	2
$DC-K-3$	M_1	M_2	M_3	M_4	M_5	M_6	M_7
S_1	13	13	13	15	15	15	15
S_2	0	0	0	0	0	0	0.9,0.9,0.9,0.9
S_3	0	0	0	0	0	0	1.99

In table A.1 DC-Knowledge, as an illustration, DC–K–1, DC–K–2, AND DC–K–3. Every row displays the transactions according to a particular category. We make the assumption that every transaction made on the user's account is known. The table's value, which shows the customers' transaction amounts (in euros) for various sequences, is used to identify patterns (importance and effect).

However, unusual evidence presents a remarkable exception. Essentially, exceptional evidence is a feature that, while it may not have been considered relevant in the training set classifications (thus having low importance), becomes significant and has a large effect in the current classification process.

Remember that this evidence model only works with statistical classifiers and that to prove effect, a second classification phase must be conducted after the training phase to ascertain importance. As such, if one were to apply this model in alternative contexts or domains, it would require redefining these two notions accordingly.

In GEN-DS, we address this challenge by reevaluating these concepts within relational knowledge structures commonly found across various applied fields.

Our original contribution lies in prioritizing specific transactions based on evidence. The significance of a transaction reflects its past relevance, while the impact evaluates its current relevance. By considering both aspects, We identify the transaction evidence or the narrative role of a transaction.

While the effect assesses a transaction's real impact, the importance metric sets the "expectation" for it. If the effect deviates from the expected importance, it becomes a noteworthy "surprise". Consequently, when a transaction has low importance but high effect, it assumes the narrative role of exceptional evidence. Normal evidence should not be highlighted in generated messages because it is not shocking and may not be worthy of attention, according to a fundamental tenet of GEN-DS. On the other hand, instances where exceptional evidence occurs are considered extraordinary and should receive prominent attention.

It is significant to remember that time and monetary worth are the two most critical variables in this specific situation. Therefore, we implement our hypotheses by asserting that: a) the significance of a telecom company's customer service may be connected to the money users spend on those services, and b) its influence can be connected to the amount users have spent on those services in the previous month [19].

A transaction can be defined as a transfer of money between a customer and the company, typically in exchange for a specific service. Every transaction sequence shows the different amounts paid for a specific kind of service over time. Consequently, the average value of normalized transactions over the previous K months is how we determine the significance of a transaction sequence. We take into consideration a time frame of the preceding six months ($K = 6$) in the following cases. Two factors influenced the choice of this duration: first, consumers would receive long messages if they went too far back in time; second, a corpus study shows that the majority of user queries relate to current transactions. The normalized value of transactions in the current month ($(K + 1)^{\text{th}}$ month) is how we define the impact of a

transaction sequence.

To normalize a transaction, divide the total amount by the largest amount the user has ever paid for that specific type of transaction. In mathematical terms, we can describe Importance and Effect as (1) and (2) if S_i represents the transaction sequence, M_j designates distinct months, and T_{ij} refers to transactions inside sequence S_i that occur during month M_j.

$$\text{Importance}(S_i) = \frac{1}{K} \sum_{j=1}^{K} \frac{T_{ij}}{\max\limits_{j=1,\ldots,K+1} T_{ij}} \tag{A.1}$$

$$\text{Effect}(S_i) = \sum_{s \in T_{iK+1}} \frac{s}{\max\limits_{j=1,\ldots,K+1} T_{ij}} \tag{A.2}$$

These numerical real values must be discretized in order to categorize their significance and impact. As per the initial model, we ascertain the smallest subset H of transaction sequences in which the overall importance/effect is at least a fraction t of the sum of the importance/effect values. When there are several subgroups with identically small amounts, we take them into account collectively. Remember that the value of t needs to be verified empirically for the particular domain. We will now give three examples of domain-specific knowledge using a threshold value of t = 75% in the following phases, as shown in [29], to exemplify these principles.

Example DC-K-1: There are three transaction sequences for the first DC-knowledge in Table A.1: S_1, for 9.99 euros (M_1–M_7), S_2, for 2 euros (M_5–M_7, appearing twice in M_7), and S_3, for 1.59 euros (M_7). We determine the significance and impact of S_1, S_2, and S_3, as well as their narrative roles, based on this data. $\text{Importance}(S_1) = \frac{1}{6}\left(\frac{9.99+9.99+9.99+9.99}{9.99}\right) = 1$ represents the significance of S_1. $\text{Importance}(S_2) = \frac{1}{6}(2+2)/2 = 0.33$ is the importance of S_2. $\text{Importance}(S_3) = \frac{1}{6}(0)/1.59 = 0$ represents the importance of S_3. Thus, the important values add up to 1.33, and the 75% equals 1. S_1 has high importance, but S_2 and S_3 have low importance. This is because $H_I = \{S_1\}$ is the smallest subset H_I such that the total of the importance values is at least 1. As a result, the effect of S_1 is $\text{Effect}(S_1) = 9.99/9.99 = 1$, the effect of S_2 is $\text{Effect}(S_2) = (2+2)/2 = 2$, and the effect of S_3 is $\text{Effect}(S_3) = 1.59/1.59 = 1$. The values in the current month determine the effect of a transaction sequence.

The effect values add up to 4, and their 75% is 3. Since $H_E = \{S_1, S_2, S_3\}$ is the smallest subset H_E such that the total of the effects is at least 3, S_1,

S_2, and S_3 all have large effects. Because of this, S_1 represents standard evidence and S_2 and S_3 represent extraordinary evidence when the discrete values of relevance and effect are combined.

Example DC-K-2: Two transaction sequences are included in this DC-knowledge example (the second example in Table A.1): S_1, which has an amount of 10 euros (M_1–M_7), and S_2, which has an amount of 2 euros (M_6–M_7). We also compute the importance and effect for S_1 and S_2 using this data. Importance(S_1) $= \frac{1}{6}\left(\frac{10+10+10+10+10}{10}\right) = 1$ is the importance of S_1. Importance(S_2) $= \frac{1}{6}\left(\frac{2}{2}\right) = 0.17$ is the importance of S_2. Hence, the importance values add up to 1.17, and their 75% is equal to 0.88. This means that S_1 has high importance and S_2 has low importance since $H_I = \{S_1\}$ is the smallest subset H_I such that the total of the importance values is at least 0.88. Effect(S_1) $= \frac{10}{10} = 1$ is the effect of S_1, and Effect(S_2) $= \frac{2}{2} = 1$ is the effect of S_2. The effect values add up to 2, and its 75% is 1.5. Since $H_E = \{S_1, S_2\}$ is the smallest subset H_E such that the total of the effects is at least 1.5, S_1 and S_2 have strong effects. As a result, S_1 represents standard evidence and S_2 represents extraordinary evidence when the discrete values of relevance and effect are combined.

Example DC-K-3: Three transaction sequences are involved in this DC-knowledge example (see the third example in Table A.1): S_1, which has amounts of 13 euros (M_1–M_3) and 15 euros (M_4–M_7); S_2, which has amounts of 0.9 euros (four times in M_7); and S_3, which has amounts of 1.99 euros (in M_7). Importance(S_1) $= \frac{1}{6}\left(\frac{13+13+13+15+15}{15}\right) = 0.94$ represents the importance of S_1, whereas Importance(S_2) = Importance(S_3) = 0. The important values add up to 0.94, and their 75% is equal to 0.71. Since $H_I = \{S_1\}$ is the smallest subset H_I such that the total of the importance values is at least 0.71, S_1 has high importance and S_2 and S_3 have low importance. Effect(S_2) $= \frac{0.9+0.9+0.9}{0.9} = 4$ is the effect of S_2, but the effects of S_1 and S_3 are both 1. The effect values add up to 6, and their 75% is equal to 4.5. One of two possible subsets $H_E = \{S_1, S_2\}$ or $H_E = \{S_2, S_3\}$ can be the lowest such that the sum of the effects is at least 4.5. Since $H_E = \{S_1, S_2, S_3\}$ is the union of the two cases, S_1, S_2, and S_3 have a high contribution. As a result, S_2 and S_3 are exceptional evidence, but S_1 is standard evidence.

We explain in the next section how the GEN-DS content selection process can be guided by our reformulation of the evidence.

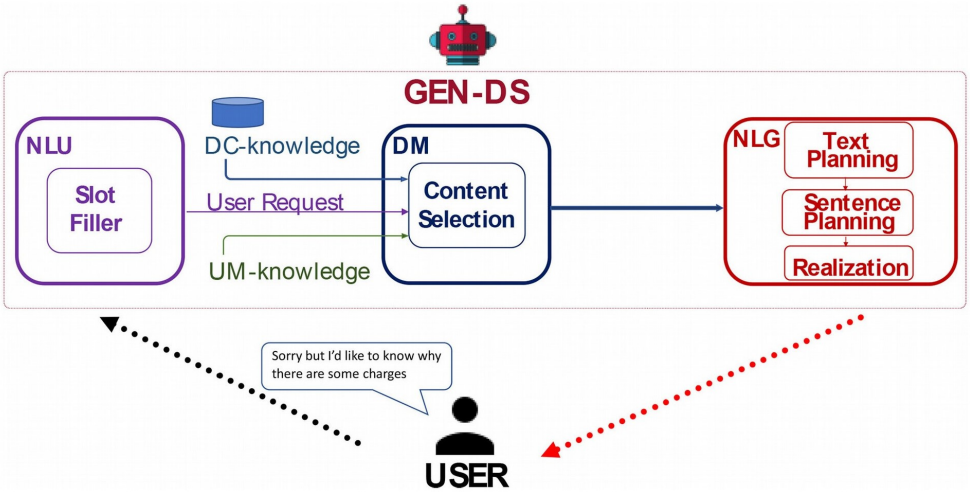


Figure A.2: The GEN-DS architecture.

A.2.3 Designing and Implementing GEN-DS

GEN-DS adheres to the traditional cascade architecture shown in Figure A.2 [116]. Natural language generation, dialogue manager, and natural language understanding comprise the GEN-DS system. Interpreting the user's statements is the NLU module's goal. The last response for the DS is produced by the NLG module. All semantic and pragmatic components that influence the future development of dialogue are managed by the DM, which processes input from the NLU module and generates output for the NLG module. For instance, when asking a question, it might choose to answer with one. Despite the lengthy history of classical architecture, new developments have been made in this field. For example, many contemporary dialogue systems utilize machine learning-based techniques in natural language understanding to populate significant conceptual spaces such as intents and entities [101], [134]).

In addition, advances in neural natural language generation can be applied in certain instances of dialogue generation within the domain [189].

Nevertheless, in task-oriented dialogue systems, the dialogue manager plays a crucial role in coordinating all the information to refresh the system's internal status and generate subsequent dialogue acts. This is not only limited to chit-chat systems but also applies to modern task-oriented DSs where effective coordination by the DM is essential [116].

As seen in Figure A.2 several task-oriented dialogue systems assume that

significant information can be extracted from the user's speech by the NLU module and then forwarded to the DM as a user request [116]. However, in certain domains such as customer care, the assumption that a highly advanced NLU module can provide detailed analysis is only partially true. This becomes evident when dealing with vague requests like the one mentioned in Example 5. (see Section A.2.1).

In this scenario, commercial dialogue systems often apologize and request users to provide more details in their requests. Additionally, the algorithm finds it challenging to assess some user utterances because they are grammatically incorrect, such as Example 3. When dealing with imprecise or grammatically incorrect user requests, GEN-DS leverages domain context and user model information to enhance response quality. The GEN-DS system, depicted in Figure A.2, uses an NLG technique that makes use of DC-knowledge to overcome the drawbacks of the apologize-and-ask-to-repeat tactic. This strategy is in line with comparable systems created for other fields [158], in the GEN-DS system, the dialogue management module plays a crucial role in generating responses that have strong support from specific domain knowledge. This allows us to anticipate user intentions and provide informative and concise answers even when faced with vague or ungrammatical requests. The GEN-DS modules for NLU, DM, and NLG are made especially to deal with explanation requests in the dataset but can also be integrated into more general dialogue systems for various types of interactions. It is noteworthy that although UM-knowledge is not employed in the process of selecting material, it does have an impact on the identification of linguistic nuances in sentences. The main characteristics of the GEN-DS NLU, DM, and NLG modules will be covered in this section.

The Natural Language Understanding module draws inspiration from the NLU features of the COM-DS system and is constructed using regular expressions. The NLU module can discern between requests for explanations that are charge-related or not. It attempts to occupy a semantic slot appropriate for the kind of request being made. For example, the NLU module returns a general user request in Example 3 while processing grammatically incorrect utterances, but returns a specific user request in Example 2.

DM Module: It deals with the process of content selection that puts into practice the concept of evidence defined in the telecom company's customer-care domain (see Section A.2.2). It accomplishes this by feeding the NLG module with all transactions together with their associated values of evidence. Furthermore, the NLG module receives information on the user

balance and total charges in certain particular cases.

For example, in Table A.1, Example DC-K-1, the content selection will convey the total charges, the transactions S_1, S_2, and S_3, along with the information that S_1 is normal evidence and S_2 and S_3 are both exceptional evidence situations.

Text preparation, sentence planning, and realization are the three essential phases of symbolic approaches to NLG that are handled by the submodules that make up the NLG module [66].

In the context of natural language generation, text planning involves determining which important information to include and organizing it in a coherent structure that follows a logical and chronological order. The DM's information is arranged in GEN-DS in part through text planning. For the sake of sentence realization, we have created a straightforward schema for organizing content into an ordered list: 1) Information about the user's balance; 2) information about the overall charges; 3) transactions containing extraordinary evidence; and 4) transactions containing normal evidence. The GEN-DS sentence planner module employs rules to ascertain the quantity and categories (e.g., declarative or passive) of sentences in the final message, selects the sentences that require merging for fluency, and chooses the lexical parts that belong in each sentence. Applications ranging from data-to-text generating initiatives have successfully used this methodology [115].

Using pre-established syntactic templates, the sentences' syntactic details are stored. The initial sentence in Example 8 is produced using the template displayed in Figure A.3. In this template, relations from dependency theory—like subject and object, known as subj and obj—are combined with phrases from constituency theory, such as Noun Phrase, NP, and Verbal Phrase, VP. It is noteworthy that the word order and word inflection are not entirely specified by these trees. Figure A.3 shows leaves that correspond to lexical items that will be defined using a realizer domain dictionary and numeric values during realization. The sentence planner adheres to two rules when selecting which templates to utilize. Since the sentences will be viewed in a text-based conversation environment where shorter sentences are preferable, the first principle gives priority to visual readability. The subject of linguistic fluency is the following premise. Combining data from transactions with the same evidence value is advised. For instance, S_1 would be communicated as a single sentence in Example DC-K-1 in Table A.1, whereas S_2 and S_3 would be integrated into another single sentence.

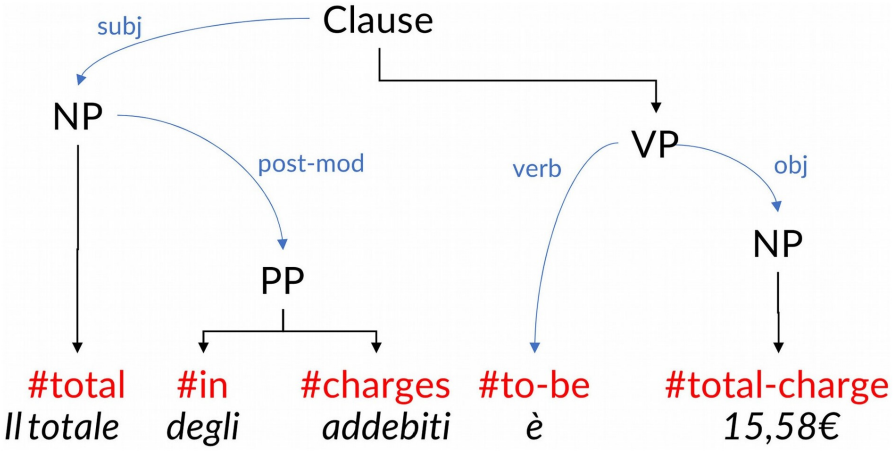


Figure A.3: Declarative sentence's syntactic template. Lexical elements that will be instantiated by the realizer are contained in the tree's leaves (shown in red, beginning with ##).

Furthermore, the execution procedure is carried out by the SimpleNLG-IT library [115], which furnishes the syntactic templates with the required orthographic and morpho-syntactic knowledge of the Italian language together with the accurate numerical values. SimpleNLG-IT is a rule-based realizer that generates morphologically correct word ordering and inflections, thereby formalizing the Italian grammar. SimpleNLG-IT is the only module in the current GEN-DS implementation that makes use of data from the user model: the Italian pronoun *tu*, which is used for young users (less than twenty years old), is a colloquial second person pronoun, whereas *lei*, which is used for older users, is a more formal second person pronoun.

In considering Example DC-K-1 as presented in Table A.1, the result that SimpleNLG-IT ultimately created is the one displayed in Example 8:

8)*Il totale degli addebiti è 15,58€. Recentemente hai pagato 4,00€ (2x€2,00) per l'Offerta Base Mobile e 1,59 € per l'Opzione ChiChiama e Richiama. Infine, come al solito, hai pagato il rinnovo dell'Offerta 20 GB Mobile (€9,99).* (The total charge is €15.58. Recently, you paid €4.00 (2x€2.00) for the Basic Mobile Plan and €1.59 for the Who Called and Call Me Back Options. In addition, as usual, you paid for the renewal of the 20 GB Mobile Plan (€9.99)).

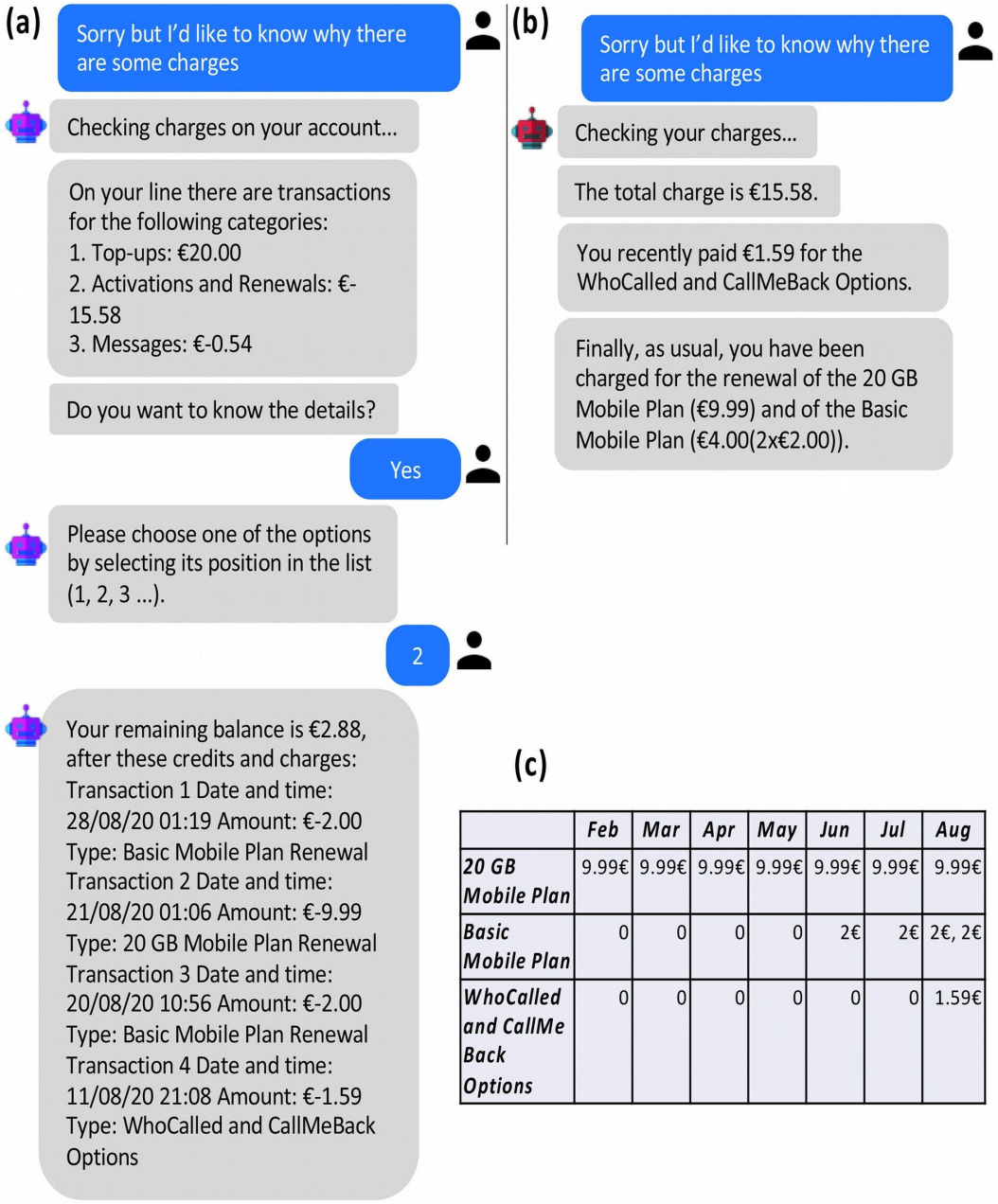
A.2.4 Building Experimental Scenarios

In this and the next section, we present the first experimental evaluation of GEN-DS carried out on human participants. Our primary objective was to create a practical evaluation for GEN-DS. Various techniques and models have been suggested in previous years for assessing DSs [44, 170, 49]. In our analysis of the sample corpus, it was found that only a small percentage (5%, as mentioned in Section A.2.1) of situations where customers require assistance involve requests for explanations, which is the specific area we are investigating. Therefore, gathering a sufficient number of samples through unrestricted live user experiments would be challenging. Thus, we have chosen to adopt one of the predefined experimental protocols [47], We generated six typical dialogues and contexts, referred to as scenarios, offline. These scenarios were then evaluated by users based on various criteria. It is significant to note that all potential customer requests can be divided into three main categories according to our study of the corpus (see Section A.2.1). These three criteria allowed us to create four different situations. Every scenario embodies a standard scenario that includes a request for clarification from the user and particular DC-knowledge. It is noteworthy that the four scenarios we developed provide strong statistical power and encompass all categories found in the corpus (see Section A.2.6). Additionally, the number of scenarios utilized in our study is comparable to previous studies [47].

Using linguistic data from Category 1, the largest category, we generated two situations (Scenario 1 and Scenario 3). Furthermore, we created two scenarios: one based on information from Category 2 and the other on input from Category 3. We chose a dialogue at random from the corpus that fit into the relevant category for each case.

We used the GEN-DS system in conjunction with the DC-knowledge linked to the dialogue and the user explanation request that started it to generate an answer for each dialogue. The translated dialogues can be found in Figure A.4

The specific DC-knowledge relevant to each dialogue was obtained from our commercial system database, which contains user transactions spanning over a period of two months. As described in Section A.2.2 of this paper, over seven months, we codified the linkages between evidence, effect, and importance. We added realistic knowledge from the last five months to supplement the DC-knowledge already obtained from the commercial system. We created Scenarios 1 and 3 to evaluate how well our formalization



	Feb	Mar	Apr	May	Jun	Jul	Aug
20 GB Mobile Plan	9.99€	9.99€	9.99€	9.99€	9.99€	9.99€	9.99€
Basic Mobile Plan	0	0	0	0	2€	2€	2€, 2€
WhoCalled and CallMe Back Options	0	0	0	0	0	0	1.59€

Figure A.4: English translation of Scenario 1 from the experiment. (a) An original conversation between a user and COM-DS that was chosen from the corpus. (b) A conversation produced using GEN-DS. (c) Common DC-knowledge.

of DC-knowledge performs in similar but different situations while maintaining the same user's request. The conversations produced by COM-DS include an average of six turns each discussion, with each turn consisting of roughly six tokens for customers and 46 tokens for agents. On the other hand, GEN-DS produces responses with only two turns per dialogue, yet each customer turn contains approximately 55 tokens while agent turns comprise about 15 tokens on average. This reduction in number of turns is due to GEN-DS promptly providing users with necessary information right after receiving a request for explanation. The dialogue corpus used in this study includes various scenarios that cover different categories and provide statistical power.

A.2.5 Participants and Experimental Procedure

To confirm the hypothesis of the experiment that users have a preference for dialogue systems that generate answers based on transaction evidence, an online questionnaire was prepared. This approach aligns with prior research conducted on similar tasks (such as [101], [189], [44], [47]). To compare the two dialogue systems, a pairwise comparison was conducted. The dialogue corpus was created using the original commercial system, COM-DS (see Section A.2.1), and an implementation of GEN-DS (see Section A.2.3) was given to the participants to assess.

We sent out emails to a number of coworkers, students, and acquaintances, inviting them to participate amiably and without compensation. We invited about a hundred people, and 54 users took part in our study. Among the total users, thirty (55.6%) were students, twenty-three (42.6%) were workers, and one (1.8%) was an educator. 29 users, or 53.7%) were between the ages of 18 and 30, 11 users (20.4%) were between the ages of 31 and 45, 13 users (24.1%) were between the ages of 46 and 60, and just one user (1.8%) was under the age of 18.

Every study participant spoke Italian as their first language, with some having no prior experience with DSs. Before participating, users were informed about the focus of the survey and reassured that no sensitive data would be collected. Users were asked to evaluate four distinct scenarios' worth of dialogue pairs produced by different customer care department specialists (DSs) in the questionnaire. Each pair of dialogues was labeled as System A and System B for ease of comparison. We deployed the questionnaire as an online survey using Google Form. There were 43 overall items in the survey: 36 of them dealt with various interaction scenarios, 6

asked questions about the user's profile (including age, education, job, and technological skills, as well as previous experience with chatbots), and 1 was open-ended for free remarks.

We followed Demberg et al.'s experimental protocol [47]: We gave users access to both the actual dialogue taken from the corpus and a dialogue produced by GEN-DS for every scenario during the questionnaire. The user explanation requests and DC-knowledge were the same in both chats. Users were asked to rank the four distinct qualities of the DSs—usefulness, necessity, understandability, and quickness—for every discussion. A seven-point Likert scale, with 1 denoting "I completely disagree" and 7 denoting "I completely agree" was used to ask users to rate their agreement with statements about these attributes. The supplemental material contains the entire questionnaire. The following are the relevant statements:

Usefulness: "All the information provided by the system is USEFUL to respond to your request" (Italian: "Le informazioni fornite dal sistema sono tutte UTILI per rispondere alla tua richiesta"). This statement's objective is to evaluate how well the conversation system's information relates to satisfying the explanation request. The user should evaluate whether all sentences from the dialogue and the associated transaction table align with the specific knowledge related to that scenario, known as DC-knowledge. Evaluating usefulness involves considering precision, similar to how it is used in information extraction.

Necessity: "All the information NECESSARY to answer your request has been presented by the system" (Italian: "Tutte le informazioni NECESSARIE per rispondere alla tua richiesta soon state presentate dal sistema"). In order to ensure that there is no extraneous information pertaining to the explanation request, the goal of this statement is to ascertain whether the information supplied in the particular scenario DC-knowledge by the DS addresses the request for an explanation. Similar to recall used in information extraction, necessity here requires considering both the sentences from DS and the specific scenario DC-knowledge within the dialogue.

Understandability: "The system provided the information in a way that it is easy to understand" (Italian: "Il sistema ha fornito le informazioni in un modo facile da comprendere"). This statement is meant to evaluate how well the dialogue system communicates during the exchange in terms of language clarity.

Quickness: "The system was quick in allowing you to find the salient information" (Italian: "Il sistema è stato rapido nel permetterti di trovare

le informazioni salienti"). This statement aims to ask users to evaluate how well the dialogue system's text generation meets the user's request for information in a timely manner. It should be noted that this factor does not pertain to the computational performance, such as response time, but rather focuses on textual characteristics like conciseness. This statement is included as there seems to be a significant correlation between user satisfaction and these specific aspects in dialogue system interactions [47].

Satisfaction: "Which system would you recommend to a friend?" (Italian: "Quale dei due sistemi consiglieresti ad un amico?"). We include a binary question in the questionnaire asking the user if they prefer one system over the other, in accordance with the evaluation schema suggested by Demberg et al. [47], in order to measure user satisfaction.

A.2.6 Experimental Results

We go over the findings for each of the previously listed properties (see Fig. A.5), which contains the 54 user's responses to the study questionnaire.

Usefulness: This statement assessed the user's confidence in the understanding of all the data the systems had presented. A two-tailed paired t-test used in the study revealed that the average for the GEN-DS system (M = 6.10, SD = 0.95) was greater than that of the COM-DS system (M = 5.35, SD = 1.58). This difference was statistically significant (t = 6.16, $p<0.001$).

Necessity: After completing the evaluation, users were asked to rate their level of confidence in its capacity to include all pertinent DC-knowledge information in the DC-knowledge . Although this difference is not statistically significant (t = 1.52, p = 0.07), the data indicate a modest preference for GEN-DS (M = 5.65, SD = 1.35) over COM-DS (M = 5.45, SD = 1.51).

Understandability: This statement assessed the user's degree of confidence over their understanding of all the data supplied by the systems. The GEN-DS system had a considerably higher mean score (M = 5.98, SD = 0.96) on this assessment than the COM-DS system (M = 4.58, SD = 1.68, t = 11.02, $p<0.001$ based on a two-tailed paired t-test).

Quickness: This sentence evaluated the user's perception of how quickly the system delivers information. In contrast to COM-DS (M = 4.35, SD = 1.56, t(215) = 13.04, $p<0.001$, a two-tailed paired t-test), the mean of the GENDS system was evaluated much higher (M=6.04, SD=1.04) on this statement.

Satisfaction: The GEN-DS system was clearly preferred during the

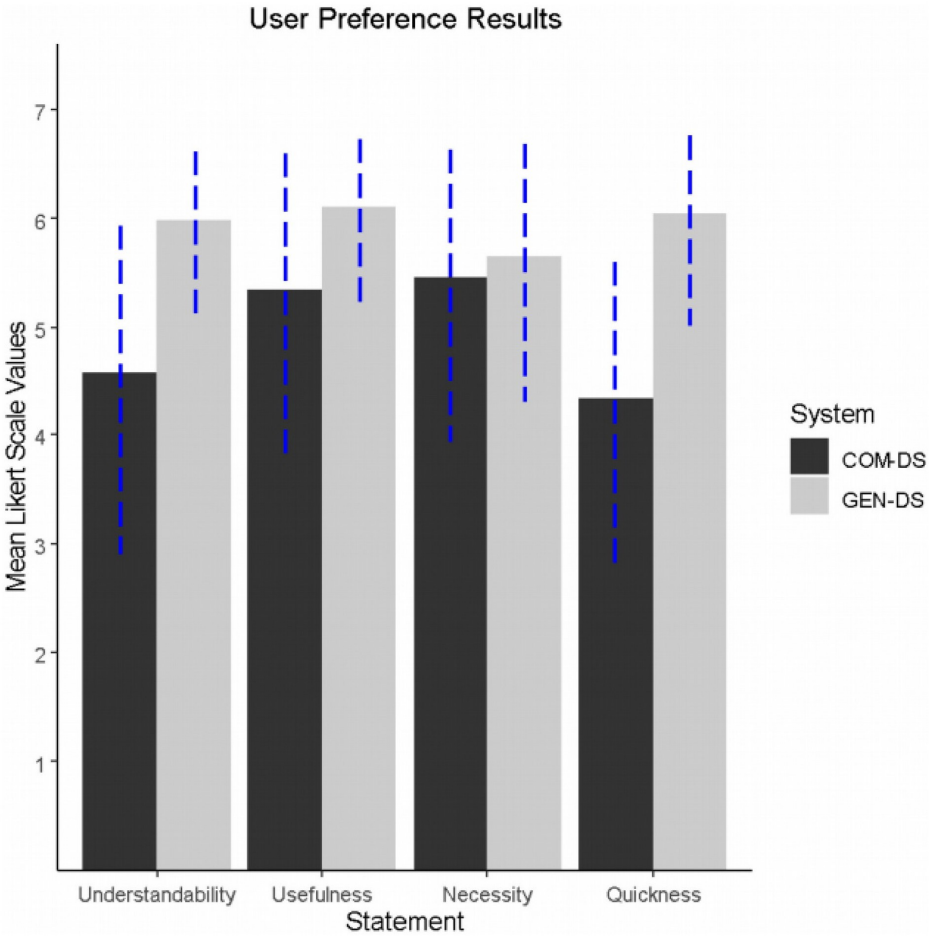


Figure A.5: The two systems' mean values and standard deviations for necessity, understandability, usefulness, and quickness.

experiment. GEN-DS was selected 157 times (72.7%) out of 216 selections (from 4 dialogue pairs and 54 participants), but the corpus-based dialogue was selected just 59 times (27.3%). A two-tailed binomial test indicates that this difference is statistically significant ($p<0.001$). The null hypothesis, according to which the corpus-based system is preferred at least as much as GEN-DS, can thus be safely rejected. We achieved high power values (>1) for usefulness, understandability, quickness, and satisfaction measurements, indicating that we had a solid sample size and enough individuals, according

to a post hoc power analysis.

Table A.2: Correlation between System Properties and User Age.

Property	**Understandability**		**Usefulness**		**Necessity**		**Quickness**	
System	**COM-DS**	**GEN-DS**	**COM-DS**	**GEN-DS**	**COM-DS**	**GEN-DS**	**COM-DS**	**GEN-DS**
Age	-.44**		-.305**		-.436**		-.371**	
	-.406**	-0.193	-.295*	-0.224	-.327*	-.344*	-0.228	-.426**

In table A.2 $*$ denotes a two-tailed significant correlation at the 0.05 level. A significant correlation at the 0.01 level (2-tailed) is shown by $**$.

A.2.7 Subgroup Analysis

In a post hoc investigation, we conducted correlation analyses to explore the relationships between user characteristics and the ratings they provided the system attributes in the survey. For numerical features, Pearson correlation coefficients were computed; for categorical features, Spearman's rank correlation coefficients were employed.

In Table A.2, you can find the correlations between users' ages and their scores for each of the four system properties. The findings showed a substantial relationship between user ratings and age. Remarkably, when combining scores from both COM-DS and GEN-DS, all four attributes showed highly significant negative correlations with age: this suggests that users, irrespective of the system, tend to offer lower scores as they age. Only a few systems/properties show significant negative correlation values when COM-DS and GEN-DS correlations are examined separately. Specifically, as age increases, the comprehensibility of COM-DS diminishes, but this does not affect on GEN-DS. However, this decrease in understandability did not have an impact on GEN-DS. For instance, individuals between 18-30 years old rated the understandability of GEN-DS at 5.35 out of 7, while those in the age range of 45-60 gave a rating of 3.98 out of 7.

In addition, older users perceive GEN-DS as being slower compared to younger users. Specifically, users between the ages of 18 and 30 rate its quickness property at an average score of 6.5 out of 7, while those aged between 45 and 60 give it an average score of 5.71 out of 7. However, these results need to be confirmed by a specific experiment meant for that purpose, as our study had a small sample size and most of the individuals were younger in age.

A.3 Discussion on Experimental Results

This part assesses our experiment's results by looking at the article's main research topic, which is whether or not users' intentions can be predicted by content selection methods, and whether or not these methods can improve conversation systems while dealing with poor language input.

Users often ask for explanations without providing enough linguistic context, according to our corpus research. Such dialogues are difficult for conventional commercial dialogue systems like COM-DS to manage well. We created GEN-DS, a system made especially to handle these difficulties, to solve this issue.The formalization of evidence enables GEN-DS to provide possible answers to ambiguous queries. Users were requested to compare, using the identical DC-knowledge and numerous attributes, the responses generated by GEN-DS and COM-DS in real conversations in a comparative experiment that is detailed in Section A.2.5.

Users believe that GEN-DS is more useful, understandable, and rapid than COM-DS, according to the data shown in Figure A.5. Users discover that GEN-DS is more efficient, useful, and clear overall than COM-DS in how it conveys the pertinent DC-knowledge. These properties are all related to how information is organized within dialogues.

The higher ratings for usefulness and quickness show that GEN-DS provides more relevant details in a clear and easy-to-understand style. The sentence planning module, which promotes shorter sentences for readability and aggregates sentences based on the evidence model outlined in Section A.2.2 to improve linguistic fluency, is responsible for these outcomes. The purpose of this study was to assess the effectiveness of various dialogue systems in customer service scenarios through an experiment.

The findings for understandability suggest that combining these principles actually enhances linguistic clarity, rather than compromising it. But in terms of necessity, there was no discernible difference between GEN-DS and COM-DS, which refers to the dialogue's ability to provide all necessary contextual information. This is expected since COM-DS already includes a comprehensive account of transactions from the past two months, exceeding the data that GEN-DS supplied. The overall user preference for GEN-DS is further supported by their satisfaction ratings when comparing both systems. In conclusion, predicting users' intentions can improve dialogue systems in cases where input quality is low.

A.4 Identified Challenges in Anticipating User Intentions

In customer care interaction systems, anticipating user intents presents several issues that, if not adequately addressed, can significantly impede the effectiveness of these systems. The critical challenges identified in our research are as follows:

Ambiguous Language Input: A pervasive obstacle we encountered was the ambiguous, vague, or ungrammatical nature of customer input. Many users communicate with dialogue systems in everyday language that often lacks clarity and structure, leading to difficulties in accurately capturing their intent.

Selective Content Presentation: Another notable challenge is determining the relevant content to present based on user queries. DSs need to extract and synthesize crucial information from a domain-specific knowledge base, presenting it concisely to address user concerns, particularly those related to unexpected charges on their accounts.

Implementing Evidence Notions: The concept of evidence in dialogue systems, while powerful, is challenging to implement. We devised a method to measure the importance and effect of transactions that impact a user's account, using these metrics to guide content selection and generate meaningful dialogue responses.

Evaluating System Performance: Assessing the performance of DSs with real users posed significant logistical considerations. Our evaluation aimed to assess whether the generated responses were useful, necessary, comprehensive, and prompt, which are attributes critically tied to the user experience.

User Demographic Diversity: Accounting for the diverse experiences and perspectives of different user demographics, such as age groups, can be exceptionally challenging but is vital in creating an inclusive DS that caters effectively to a broad customer base.

Sample Size Limitation in Experiments: The limited sample size of users available for testing and providing feedback on the DS remains a significant constraint. Broader and more diversified user participation is crucial for a more comprehensive analysis of the DS's performance across various metrics.

Addressing these challenges is imperative to enhance the anticipatory capabilities of DSs, thereby improving user satisfaction and retention, which

are paramount in customer care settings.

A.5 Future Directions

Considering how we want to go with "Anticipating User Intentions in Customer Care Dialogue Systems", we see several areas for further research and advancement. The primary focus is to improve our evidence-based response generation, expanding its application beyond customer care to explore its effectiveness in other service domains. Additionally, we plan to integrate a comprehensive user model that incorporates demographics, behavior patterns, and interaction histories to enhance system responsiveness and personalize dialogue responses.

We plan to improve the system's interactive capabilities for more dynamic exchanges, including soliciting additional information from users when faced with uncertainties. Enhancements to the user interface are also crucial, aiming to deliver a seamless and engaging user experience through multimodal communication features.

Future studies will aim to include a more diverse user demographic in our data set for a thorough assessment of the conversation system. Through rigorous testing on a larger scale, we intend to overcome current sample size limitations and enhance the generalizability of our findings. We also plan to actively gather user feedback on system performance and effectiveness, using this information to make iterative improvements and optimizations.

A.6 Conclusion

With this study, we showed how useful it is to include evidence in dialogue systems for customer care. Through an analysis of a corpus, we discovered that many user requests at the beginning of conversations with DSs were unclear or grammatically incorrect.

Consequently, dialogue managers lacked sufficient linguistic information to generate meaningful responses in these cases. To address this issue, most commercial DSs employed a simple strategy of apologizing and requesting users to repeat their queries. To overcome this issue, a possible approach is to utilize the non-linguistic knowledge associated with the conversation, such as user-related information and domain-specific knowledge.

DC-knowledge in customer service contexts refers to business-to-business exchanges between customers and businesses. Relational frameworks are

commonly utilized in the organization of this kind of data. To help choose pertinent content from DC-knowledge, we have provided a clear concept of proof for relational data. To implement this concept, we developed GEN-DS, a system that generates text based on data.

In our experiment, we contrasted synthetic dialogues produced by the GEN-DS system with real dialogues from a corpus of human/commercial dialogue system exchanges for a telecom company's customer care service. The questionnaire's results showed preferences for GEN-DS dialogues in terms of their overall enjoyment, usefulness, necessity, and understandability. This illustrates how this strategy can improve the quality of conversations in the customer-care arena.

In order to evaluate GEN-DS in a scenario where user sentences are difficult to understand, we conducted a simulated dialogue experiment instead of using real conversations. Although this type of experiment has limitations, the nature of our research objective does not permit us to engage with users in natural interactions to assess the influence of content selection determined by evidence.

Moreover, we believe that our study's conclusions can be extended to fields other than customer service. Our research is based on the formalization of relational knowledge evidence and its incorporation into natural language creation. First introduced in the field of machine learning, Biran and McKewon proposed this idea. Expanding on their definition of importance and effect as types of evidence representation, we suggest encoding system behavior history as "importance" and recent behavior as "effect". In order to apply our approach within DC-knowledge based NLG tasks, it would necessitate adapting these concepts specific to DS-related requirements.

Formalizing the notion of evidence in relational knowledge and applying it to natural language production in the customer service industry is the main goal of our research. Biran and McKewon first proposed this idea in the field of machine learning, where they characterized evidence in terms of significance and effect. Building on their work, we proposed two key elements: incorporating past system behavior into importance and recent behavior into effect. To implement our evidential approach with DC-knowledge, Evidence, importance, and effect are concepts that must be modified for the particular task related to DS.

Lastly, our research raises a significant query regarding the combination of user demands, UM-knowledge, and DC-knowledge. Future research

should investigate the role of the user model in this process and how evidence might be applied to scenarios with understandable language input.

A.7 Major Contributions in Anticipating User Intentions in Dialogue Systems

In this chapter, we have highlighted significant contributions throughout various important sections.

- We conducted experiments to test the hypothesis that users favor dialogue systems based on transaction evidence. To measure user preferences, we used an online questionnaire and compared two dialogue systems, COM-DS and GEN-DS. We employed pairwise comparison to evaluate user preferences during the assessment of a dialogue system.

- Our analysis of the user evaluation results was important in determining the significance of GEN-DS. Users rated GEN-DS higher than COM-DS in terms of usefulness, indicating their trust in the provided information's relevance. This finding showcases the effectiveness and relevance enhancements of our GEN-DS system.

- We analyzed the relationship between user characteristics and system ratings. This included calculating correlation coefficients, which showed significant negative correlations between user age and certain properties of the system. These findings highlight how demographic factors can impact preferences and effectiveness in dialogue systems.

- To compare the performance of GEN-DS with COM-DS, we performed a statistical analysis. The results of tests like the binomial and paired t-test showed that GEN-DS was much preferred over COM-DS. This analysis provided strong statistical evidence supporting our experimental hypothesis and rejecting the null hypothesis.

In our research, we made significant contributions to the field of customer care dialogue systems. Specifically, we focused on understanding user preferences and evaluating the effectiveness of the GEN-DS system in meeting these preferences. Through experiments and statistical analysis, we gained valuable insights that contribute to advancing knowledge in this area.

A.8 My personal contributions for the Anticipating User Intentions in Dialogue Systems

I played a vital role in conducting experiments to assess user perceptions across various crucial properties, including usefulness, necessity, understandability, and quickness for both dialogue systems GEN-DS and COM-DS.

Through meticulous data analysis of user responses obtained during the experiments, I calculated mean values and standard deviations using a quantitative approach that allowed us to derive precise insights into user preferences and satisfaction levels. My focus was on uncovering the complexities of customer service conversation systems by comprehending user preferences and assessing the performance of the GEN-DS system.

To compare GEN-DS and COM-DS, I utilized statistical tests such as t-tests and binomial tests. The examination revealed noteworthy differences between the two dialogue systems providing compelling evidence favoring our experimental hypothesis for GEN-DC.

Additionally, I investigated subgroup analysis like as correlations between user attributes and system ratings using correlation coefficients to measure connections. The findings demonstrated significant negative correlations between user age and specific system properties which offer insights into how factors like age can influence users' preferences thereby underscoring the significance of considering demographic elements when designing or enhancing customer care dialogue systems.

My contribution added valuable insights to our research efforts aimed at advancing knowledge in this domain.

Appendix B

Thesis Appendices

B.1 Published Article 1: Machine learning algorithms distinguish discrete digital emotional fingerprints for web pages related to back pain.

In my thesis, we explore the application of machine learning algorithms to distinguish discrete digital emotional fingerprints for web pages related to back pain, as discussed in the study by Caldo, Davide, Silvia Bologna, Luana Conte, Muhammad Saad Amin, Luca Anselma, Valerio Basile, Md Murad Hossain, and Giovanni De Nunzio (2023).

This research, published in Scientific Reports, employs innovative methodologies to analyze the emotional content of web pages using computational techniques. we highlight the significance of understanding emotional nuances in online health-related content, particularly focusing on back pain. Our findings underscore the potential of machine learning in extracting and interpreting emotional signals embedded in digital information, thereby contributing to advancements in sentiment analysis and content summarization within the medical domain.

scientific reports

Check for updates

OPEN

Machine learning algorithms distinguish discrete digital emotional fingerprints for web pages related to back pain

Davide Caldo[1,8]✉, Silvia Bologna[2,8]✉, Luana Conte[3], Muhammad Saad Amin[4], Luca Anselma[4], Valerio Basile[4], Md. Murad Hossain[4], Alessandro Mazzei[4], Paolo Heritier[5], Riccardo Ferracini[6], Elizaveta Kon[7] & Giorgio De Nunzio[3]

Back pain is the leading cause of disability worldwide. Its emergence relates not only to the musculoskeletal degeneration biological substrate but also to psychosocial factors; emotional components play a pivotal role. In modern society, people are significantly informed by the Internet; in turn, they contribute social validation to a "successful" digital information subset in a dynamic interplay. The Affective component of medical pages has not been previously investigated, a significant gap in knowledge since they represent a critical biopsychosocial feature. We tested the hypothesis that successful pages related to spine pathology embed a consistent emotional pattern, allowing discrimination from a control group. The pool of web pages related to spine or hip/knee pathology was automatically selected by relevance and popularity and submitted to automated sentiment analysis to generate emotional patterns. Machine Learning (ML) algorithms were trained to predict page original topics from patterns with binary classification. ML showed high discrimination accuracy; disgust emerged as a discriminating emotion. The findings suggest that the digital affective "successful content" (collective consciousness) integrates patients' biopsychosocial ecosystem, with potential implications for the emergence of chronic pain, and the endorsement of health-relevant specific behaviors. Awareness of such effects raises practical and ethical issues for health information providers.

Degenerative musculoskeletal chronic pain is a very significant and costly problem throughout the industrialized world with low back pain being the leading cause of disability[1]. The current concept of treatment applies when the local anatomy is macroscopically altered as demonstrated by imaging and alterations become an established source of chronic pain; treatments largely target the organic substrate and include analgesic drugs and surgical decompression or fusion[2]. Current approaches showed major limits: systematic reviews find scant evidence of medication efficacy[3] and diffuse severe complications following opioid widespread use[4].

Musculoskeletal chronic pain is also known to be related to disease-specific psychological and social components, intertwined in a variety of proportions with the organic substrate, for instance: fibromyalgia is mostly driven by central nervous predispositions[5]; hip and knee osteoarthritis have a larger peripheral nociceptor contribution, also driving a better success rate of pain relief with joint replacement surgery[6]; as high as 90% of back pain cases are classed as non-specific, as there is often no definite organic substrate cause for the experienced pain, putting it somewhere in between the fibromyalgia and the great joint degeneration end of the spectrum of musculoskeletal degenerative diseases[7]. Failed back surgery rates 20–40% resulting in persistent pain[8].

Specific emotional patterns emerged to relate to different musculoskeletal degenerative diseases[9–17]; the emotional content of external sources of information can complement rather than mirror users', in a non-straightforward complex system, in analogy with the counterintuitive role of sadness in the pleasure elicited by

[1]Humanitas Gradenigo Hospital, Turin, Italy. [2]Imparamare Ong, Asti, Italy. [3]Mathematics and Physics Department "Ennio de Giorgi", University of Salento, Lecce, Italy. [4]Informatic Department, Turin University, Turin, Italy. [5]Digspes Department, Oriental Piedmont University, Alessandria, Italy. [6]Genoa University, Genoa, Italy. [7]IRCCS Humanitas Research Hospital, Milan, Italy. [8]These authors contributed equally: Davide Caldo and Silvia Bologna. ✉email: davide.caldo@gmail.com; dr.silvia.bologna@gmail.com

sad music[18]. Received information acts as a countering or favoring agent for the emergence of chronic pain in the biopsychosocial arena[19]. Affective content plays an integrated role in the process[20].

The dominant source of information for patients in modern society is the internet[21]. The success of internet pages is the result of a dynamic interplay between users that "socially validate" selected pages (by visualization, interaction, and other actions) and search engine algorithms that amplify the visibility of such validated pages, raising them toward the top search output listed pages; further social validation is thus acquired and a phenomenon of self-increased popularity is engaged; the final results is that a small fraction of high ranked pages inform the great majority of users[22].

Sentiment analysis (also known as opinion mining) is the systematic identification, extraction, quantification, and study of affective states using natural language processing, text analysis, and computational linguistics via computer science[23].

Machine Learning (ML) can analyze large amounts of data and perform classification with measurable predictivity[24].

On such premises, we formulated the general hypothesis that "successful" internet pages related to a given pathology can trigger their own validation by users by embedding a specific, consistent emotional pattern, complementary to patients' biopsychosocial ecosystem.

Thus, supervised ML algorithms would accurately discriminate the topic of the original text by the affective emotional fingerprint produced by sentiment analysis. In the present research top ranking English language websites are analyzed, comparing two pooled conditions: the first one is the "nonspecific" degenerative chronic lumbar back pain (LBP), characterized by a cluster of multiple organic substrate alterations variably combined and highly non-specifical, leading to high rates of failed surgical treatment; the other is great joint (hip and knee) chronic degenerative disease, with more specific substrate alteration, generally leading to the higher success rate of surgical treatment.

The scope of our work is to test whether a consistent affective pattern characterizes the corpus of knowledge socially validated and related to a specific pathology, allowing discrimination from a control group characterized by a different relative specific weight of biopsychosocial components. In particular, the presence of discriminating emotions would be consistent with somatic marker theory applied to the emotional domain in medicine[20]. Such a result would suggest a specific role for emotional fingerprints. As far as the authors know there is no previous systematic analysis of the affective content of medical internet pages concerning low back pain and other musculoskeletal degenerative diseases. This is a relevant gap of knowledge since digital emotional content integrates the patient's biopsychosocial system, with a potential role in the emergence of chronic pain and enforcing/blocking positive/negative social behaviors; major relevance implications would arise for institutions responsible for the diffusion of healthcare-relevant information and in general for medical information providers on the internet both on the practical and ethical plane. If the hypothesis were verified, ethical considerations related to the largely subliminal effects of emotional elements embedded in website information would need further discussion by scientific and policy-making representatives.

No humans were involved in the study, which is based on data openly available from the authors upon reasonable request.

The paper is structured as follows. In "Literature review" we discuss the studies that constitute the background of the study. In "Results" we introduce data description and data analysis results. In "Discussion" we comment the results and provide some possible interpretations suggesting further implications. Finally, in "Materials and methods" we describe the data selection process, the identification of Internet resources, and the data analysis process.

Literature review

The notion of information being a key factor in health/disease emergence is in general consistent with the Biopsychosocial (BPS) model that suggests that a person's state of wellness or illness is not coincident with the organic substrate alteration but rather intertwined with psychological and social factors[25]. The critical role of Emotions, outlined by decades of affective neuroscience findings, tightly integrates affective domains in the model[20]. Emotions ultimately bias or determine behavior according to the somatic marker theory[26,27]. In fact, the emergence of chronic pain in neuroscientific literature has been linked via functional neuroimaging to the activation of several sensorial, cognitive, and affective central nervous system circuits[28]. The main BPS model limit remains the lack of tools for clinical usability: thus, the heuristic value of a reductionistic "mechanical" approach largely prevailed in clinical settings[29]. The BPS model evolved by incorporating notions from Ashby's law of requisite variety, Rothman's notion of multiple sufficient causes of a condition, and top-down causation in complex adaptive systems[30]. Within such a theoretical frame degenerative musculoskeletal chronic pain can be interpreted as an emergent property of a complex adaptive system, with affective information being a critical element of the system[19,30].

LBP is the main source of chronic pain and disability in the world[1], linked in a controversial way with emotional regulation, somatosensory amplification, and rumination in negative affective or dysfunctional beliefs[10]. LBP psychological arena more frequently includes fear[11], anger, and sadness in a frame of catastrophism, anxiety, or depression[12]. Fear has a direct effect on the outcome of patients, influencing behavior: fear of pain and/or injury/movement leads to movement avoidance, and it is possibly implicated in the transition from acute to chronic and the persistence of disabling LBP[31]. Anger is another leading emotion in many LBP studies, with greater effects on chronic pain severity than sadness: it is shown that people who tend to express anger and who exhibit high pain sensitivity could be characterized by deficits in endogenous inhibitory mechanisms[32]. A symptom-specific reactivity model showed that anger arousal may lead to increases in muscle tension near the

site of injury, and thereby increase pain; increases in lower paraspinal muscle tension are higher in anger than in sadness, and patients with elevated anger expressiveness showed greater increases in muscle tension[13].

The estimated 2010 prevalence of total hip and total knee replacement among the total US population was 0.83% and 1.52%, respectively[33]. Chronic pain despite joint replacement is not uncommon, affecting approximately 10% of patients after total hip replacement and 20% of patients after total knee replacement[34]. The related emotion reported in scientific literature is fear[14], altogether with withdrawal and depression[15]. Psychological and structural factors interact exacerbating pain perception[16,17].

A characteristic of people with chronic pain is avoidance: the "cognitive-behavioral fear-avoidance model" includes cognitive (idiosyncratic maladaptive beliefs on pain), affective (fear), and behavioral (avoidance) components[35].

Disgust is one of the basic emotions characterized by a strong sense of aversion, associated with physical reactions (nausea, sweating and lowering blood pressure)[36]. It is an emotion of well-rooted evolutionary origin: animals have evolved a series of behaviors to reduce the risk of infection by pathogens, such as microorganisms[37]. Disgust showed the tendency to lead to a complex regulation of immune-related functions, effects similar to the acute phase response to infections; immediately after a disgust induction, reported pain is reduced, but later it increases leading to a final higher pain sensitivity[36].

Although disgust was first thought to be a motivation for humans to avoid only physical contaminants, it has since been applied to psychosocial contaminants as well. Likewise, when a group experiences someone who commits violence or misdemeanor to another member of the group, its reaction is to "divert" that person from the group, basically the same recorded when contaminating fluids are involved[38]. When one experiences disgust, this emotion might signal that certain behaviors, objects or people are to be avoided in order to preserve psychosocial purity[39], as opposed to other emotions such as fear, anger, and sadness that appear "unrelated to moral judgments of purity". The emotion of disgust can be hypothesized to serve as an effective mechanism following occurrences of negative social value provoking repulsion and desire for social distance. The origin of disgust can be defined by motivating the avoidance of offensive behavior, and in the context of a social environment, it can become an instrument of social avoidance. Disgust is known to reduce motivations for social interaction[40]. Social interactions are a key component of well-being in the aging population, with the growth of emotional empathy serving as a compensation factor to cognitive age-related decadence of cognitive empathy[41].

Several studies have shown that the internet (through social media) can be a representative source of data, exploitable to recognize public perceptions and behaviors during a crisis, and even predict outbreaks[42,43]. In modern society, information is increasingly being sought on the internet[44]. In the medical field, its relevance has grown rapidly: in the United States, for example, the number of people seeking medical information on the Internet has increased from 54 million in 1998 to about 117 million in 2005[45]. In January 2022, 4.66 billion people accessed the internet; GWI's survey also finds that 25.9 percent of working-age internet users check health symptoms online every week[44].

ML algorithms showed predictive accuracy in many biology, medicine, and even social system applications; precision medicine in the twenty-first century strives for accurate prediction of what is beneficial for individual patients; prediction, as opposed to association, comes into play when forecasting outcomes that are yet unobserved. Nonetheless, relevance is not a synonym for discriminant power as used in classification and prediction: significant variables in a statistical model do not guarantee prediction performance, and non-significant attributes might reveal predictive[46]. Machine learning algorithms have been previously applied to the field of pain affection[47].

An entirely new discipline called Social Neuroscience is based on neurophysiological (i.e. mirror neuron and internal simulation activity) and social evidence of the notion of collective consciousness (CC), the set of shared knowledge, assumptions, moral attitudes operating as a unifying force within society[48]; CC includes collective emotions, emergent macrolevel affective processes that cannot be readily captured at the individual level[49]. The CC "storing" substrate evolved in time: by the end of the eighties of the twentieth century less than 1% of the world's information was archived in a digital format, whereas in 2007 this percentage reached 94% and rapidly growing[21]. The original concept of a Virtual Collective Consciousness derived from humanities and was initially restricted to social networking influencing behavior[50,51]. The tight relation of CC with digital information is the result of a dynamic interplay between users and search engine algorithms: users determine "social validation" of web pages by interacting with the ones best complementing their psychological drive; search engine algorithms (like Google's PageRank) further enhance visibility of those pages putting them at the top of the search engine query results, closing a self-powered circuit for the constitution of a pool of "successful" and widely accessed digital information[22]. The digital information feeds the cognitive and affective knowledge of patient collectivity: it has been proposed that multiple people can converge on the same emotion pattern when exposed to digital media[52] and that the BPS model must evolve to include a "digital" component, developing a 'biopsychosocial-digital' approach to health[53].

Results

Relevant sources identification. We a selection of an arbitrary number of 2000 sites in English, French, German, Italian and Spanish with high relevance/popularity for the conditions of interest. The sites were listed in an Excel diagram. We then identified 837 Sites in English. It was plotted the qualitative fingerprint of all the different languages; the English language pages were forwarded to quantitative analysis for greater consistency in the result, to exclude variables related to linguistic influence that may be the subject of subsequent studies. The analyzed sites concerned the following conditions of musculoskeletal pathology: back pain, herniated disc, hip prosthesis, knee prosthesis, low back pain. Within the 837 English sites the various pathologies were distributed as illustrated in Table 1.

Label	Count
Back pain	165
Disk herniation	176
Low back pain	164
Hip prosthesis	168
Knee prosthesis	164

Table 1. Number of documents considered for each pathology (English sites).

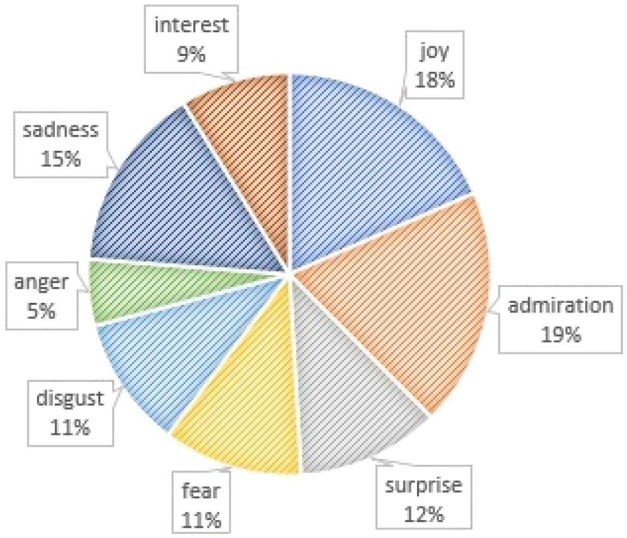


Figure 1. Prevalence of emotion words in all selected sites.

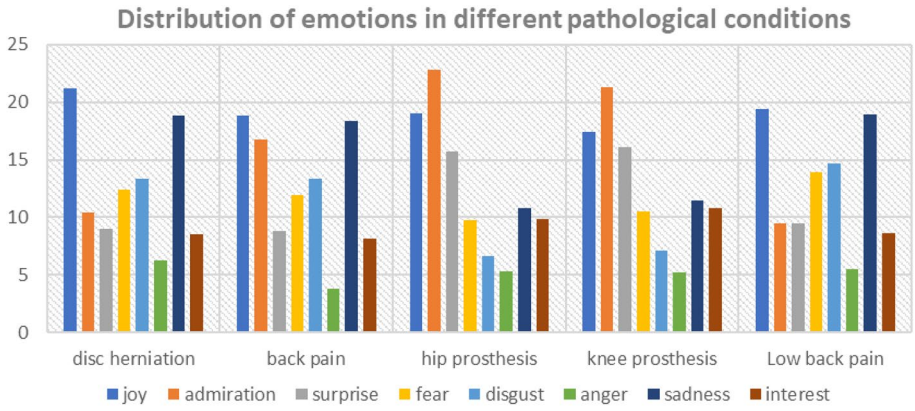


Figure 2. Charts illustrating prevalence of emotion words in different conditions.

Semantic analysis. Emotion content word distribution, normalized to number of emotion words per site, is depicted in Fig. 1. Distribution of emotion content words per condition and per emotion is illustrated in Fig. 2.

The qualitative analysis leads to a data visual pattern: the emotional fingerprints in Fig. 3. The fingerprints are substantially histograms of per-document (scaled) counts of emotional content words: each histogram occupies a column in the image, with bins arranged top to bottom and refers to the one of the considered emotions, so each fingerprint contains 8 hystograms. Warm colors indicate highly occupied bins while cold colors are empty bins.

Nonparametric statistical tests and machine learning. We decided to compare the subsets of texts for the different health conditions in a pairwise fashion, both graphically with scatterplots, and computationally with significance tests (Mann–Whitney U-test) and a ML approach, to assess if the sets of documents significantly differed at group level and at individual level.

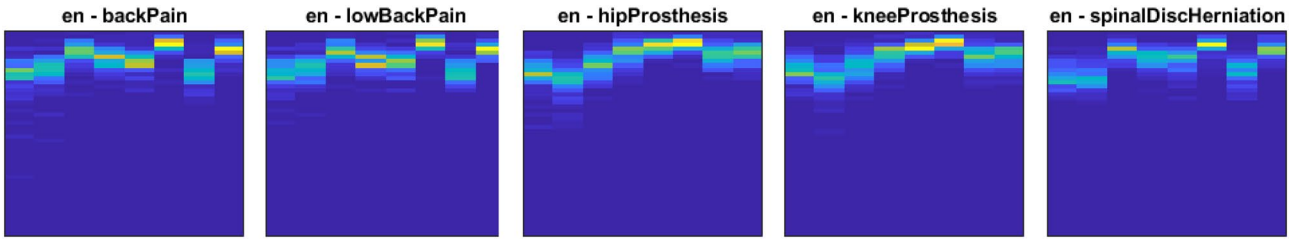


Figure 3. Emotional fingerprints for the documents related to the five health conditions considered (English language). In the fingerprint, eight "pixel" columns left to right refer to joy, admiration, etc., and the intensity of the emotion (in arbitrary scale) increases top to bottom.

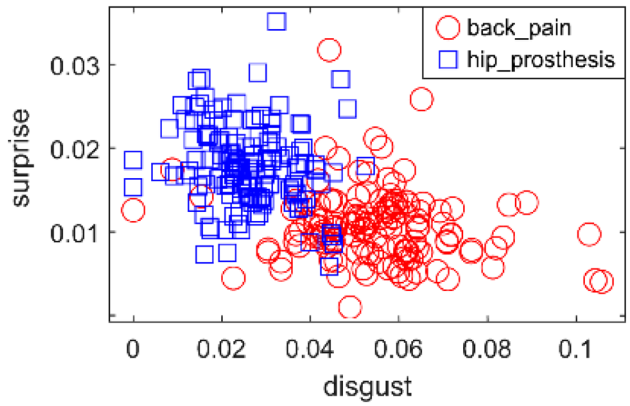


Figure 4. Scatterplot showing the distributions of the Emotional score variables related to disgust and surprise, for back pain vs hip prosthesis documents (English language).

The internet documents were modeled as vectors of variables (the emotional scores for joy, admiration, surprise, fear, disgust, anger, sadness, and interest) and labeled by the health conditions of interest ("classes").

Some variables (in particular, disgust) showed a large discriminating power. Some pairs of variables, too, were discriminating when considered together (e.g., disgust and surprise). A few features were quite strongly correlated, such as disgust and sadness. Figure 4 shows the example case of scatterplot for English language, back pain vs hip prosthesis documents. In this example, surprise and disgust are the two paired emotions and their peculiar distributions are signals of important discriminating power.

Statistical nonparametric tests (Mann–Whitney U-test) applied to assess univariate emotional score group differences between documents for different health conditions, mostly failed in finding low p-values (which would allow to reject the null hypothesis that the documents from two classes come from the same distribution). On the contrary, ML employed to classify the documents detected many cases with large classification accuracy, which proves that document emotional contents have peculiar patterns in each class of a pair of health conditions, making them different and recognizable. The accuracy and other statistics for the English-language documents were calculated for all the pairs of conditions of interest, as shown in Table 2 which reports the interesting cases with accuracy > 0.9 for linear Support Vector Machine (SVM) classification.

Using decision trees instead of SVMs gave very similar results and allowed us to explore the role of the emotion variables. The graphics in Fig. 5 show the decision tree before and after pruning. Pruning is a tree compression technique that removes sections of the decision tree that are redundant and have little influence on classification accuracy. Pruning decreases classifier complexity and has two consequences: it reduces overfitting (possibly enhancing generalization) and improves explainability. Figure 6 shows the estimates of predictor importance, consistent with the trees.

Discussion

To give back depth to pain experience in musculoskeletal degenerative pathology it is necessary to reinterpret it within the BPS approach and consider its information source feeds with specific attention to the emotional component. Digitalization shifted from an irrelevant share to nearly the totality of the health knowledge corpus, coupled with widespread access to the internet, dramatically changing the scenario of healthcare information. The described reciprocal interaction between the users and the affective content of digital medical information selects a highly representative sub-corpus of information: for practical reasons, it is proposed to be referred to as "digital affective collective consciousness" (DACC). The DACC includes the affective subset of the "virtual collective consciousness", embedded in successful pages related to every area of human knowledge. Sentiment analysis is a useful method to characterize DACC projection to entire sectors of information or specific topics

Support vector machine classifier	Confusion matrix	Recall or sensitivity	Specificity	Precision	Accuracy	F1 Score	Mann–Whitney U-test p-value
Back pain vs hip prosthesis	[159, 6; 8, 160]	0.95	0.96	0.96	0.96	0.96	0.48
Back pain vs knee prosthesis	[160, 5; 8, 156]	0.95	0.97	0.97	0.96	0.96	0.049
Disc herniation vs knee prosthesis	[165, 11; 10, 154]	0.94	0.93	0.94	0.94	0.94	0.96
Disc herniation vs hip prosthesis	[162, 14; 11, 157]	0.94	0.92	0.92	0.93	0.93	0.062
LBP vs hip prosthesis	[156, 12; 10, 154]	0.94	0.93	0.93	0.93	0.93	1.2
LBP vs knee prosthesis	[155, 9; 6, 158]	0.96	0.95	0.95	0.95	0.95	0.28

Table 2. The columns report various classification statistics of a linear SVM classifier trained and validated with a fivefold cross-validation scheme for the English language and for the discrimination between various health conditions in pairs (cases with accuracy > 0.9 were selected); 2nd to 7th columns contain the confusion matrix [*true positives, false positives; false negatives, true negatives*], sensitivity, specificity, precision, accuracy, and the F1 score. The last column shows the p-values issued by the Mann–Whitney U-test.

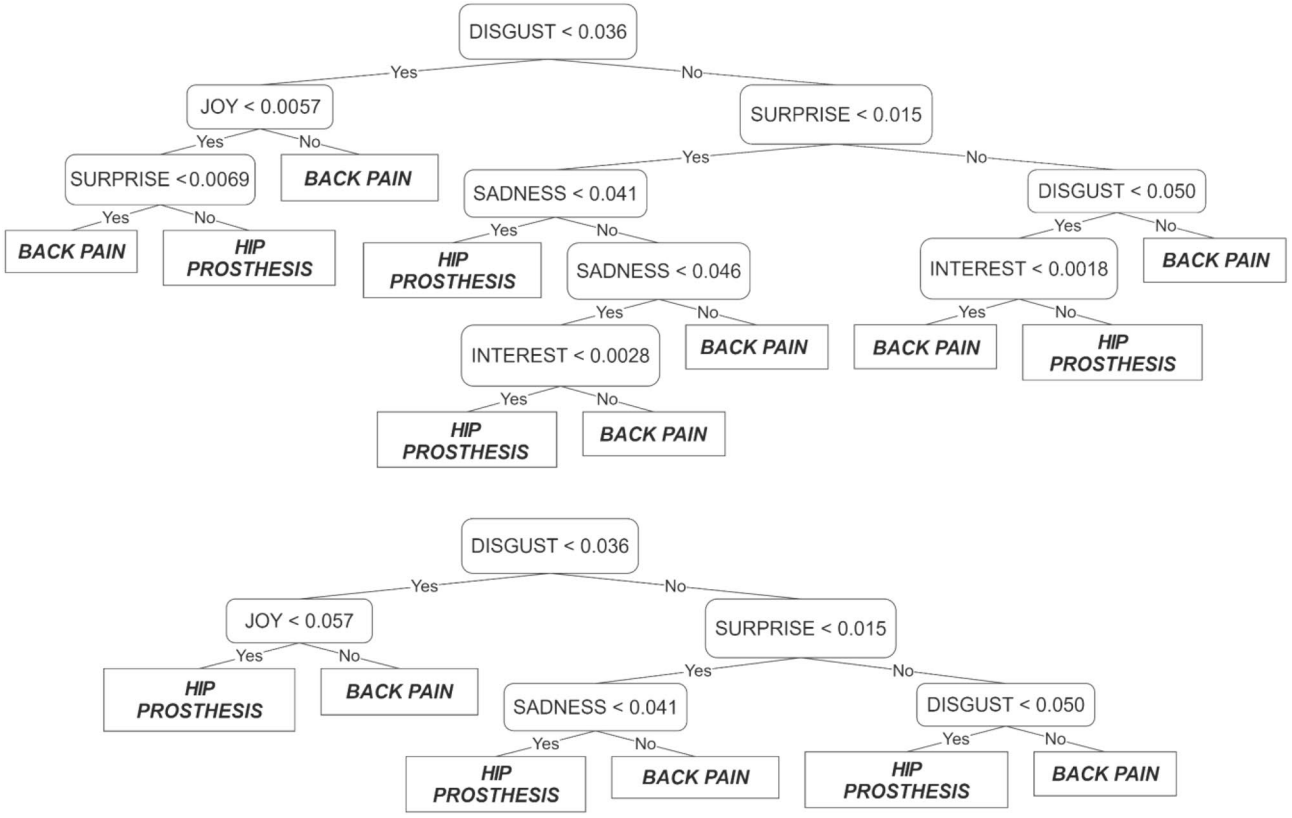


Figure 5. Decision tree before and after pruning, for the English language, back pain vs hip prosthesis.

on the internet, relating to a basic emotion model; such assessment is extremely relevant when pathologies with a major psychosocial content are investigated. LBP generally does not relate to a specific major organic lesion, but rather the result of a combination of many minor degenerative lesions with a "non-specificity" that parallels fibromyalgia, and a major psychosocial component, including a vast array of emotions. Emotions ultimately bias or determine behavior according to the somatic marker theory according to somatic marker theory[26,27].

We identified substantial histograms of per-document (scaled) emotional word counts as an easy and novel graphical way to plot quantitative presence of emotion-related words with their intensity in the emotional scale.

From our digital information analysis, major relevance for disgust emerged as a key discriminating factor between LBP and hip/knee affection-related internet pages; disgust acts as the first bifurcations in all the ML decision trees generated (see "Materials and methods" and "Results"); in some cases, the degree of disgust alone identifies the original topic from the affective pattern; in other cases, ML relies on a combination of disgust with

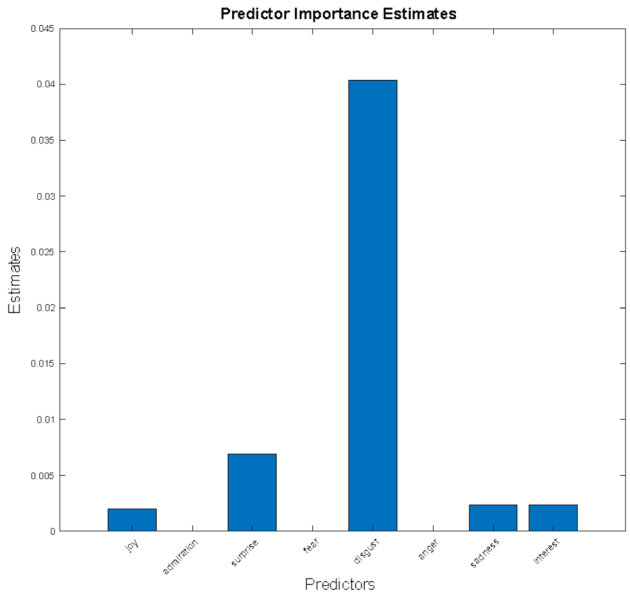


Figure 6. Estimate of predictor importance, for English language, back pain vs hip prosthesis.

surprise, joy or sadness. The difference in disgust intensity in the two affective patterns of spine and control pathology may reflect the different biopsychosocial profile of affections.

Disgust is a diverging emotion: it would not be expected to be associated with popularity. It can be argued that digital information of disgust could be involved in the "adaptive" characteristic of the pain system, the "diverting from disease" impulse endorsing avoidance strategy. It can be conjectured that the temporary decrease in pain sensitivity plays a role in internet pages' success. If that is the case, the later sensitization to chronic pain may contribute to a net negative outcome. Such a scenario calls for specific research since it raises serious concerns about digital information subliminal emotional long-term effects; scientific, but also ethical, and legal implications are raised.

In general, the role of disgust may relate to different ways of perceiving some characteristics intrinsic to each of the two groups of affections.

Differences in the emotional profile of the patients have been addressed in the Literary Review section. Furthermore, spine and great joint degenerative conditions both affect walking capability, but with different pathways[54]. This may be a potential source of the psychosocial divide between the two groups of conditions, reflected in the pages promoted in the internet user-algorithm dynamic; in the structure of disgust motion is less relevant compared to other emotions such as fear, since disgust can relate to disembodied entities, such as moral judgment.

The component of attraction power detailed by Sachs[18] is based on the pure aesthetic power of the art; in that sense diverting emotions such as sadness or disgust could be felt on its aesthetic value alone. The authors are not aware of any published study concerning the assessment of the intrinsic aesthetic feature and role of medical websites, in particular concerning pages related to spine, hip, or knee pathologies.

Some limits of the present work are acknowledged. The control group was chosen arbitrarily; the decision, though, was based on a similar condition from a medical perspective (concerning epidemiology, timing, and roadmap of treatment) to elude potential hidden variables, but with the characterizing differences of the surgical outcome. In our results statistical association did not match predictivity; it has been shown the latter represents the best approach to complement explanatory models and setting new neuroscientific theoretical grounding[24], in our work we combined the explanatory power of the somatic marker theory with the predictive power of modern computational techniques to overcome limitations of mere statistical association, namely ML algorithms to predict the pathology from the emotional content of the page. Furthermore, we extended the implications of somatic market theory to social neuroscience in the arena of the digital information society. This could pave the road not only to a deeper understanding of biopsychosocial pathologies but also to an updated version of the bps model. Also, we point out that the fingerprint methodology is a novel representational construct of emotion representation, particularly fitting for comparative tests (in our case, an intra-linguistic one).

More tests need to be performed to fully outline the extent and generalize the relevance of the findings, laying the ground for subsequent potential applications. Quantitative analysis was limited to the English language only to contain language-dependent biases. The study of fingerprints (Fig. 3) may be taken up in later studies for inter-language comparison. Only the presence of words attributed to specific emotions is considered by Senticnet; more advanced methods also consider text complexities such as negation or sarcasm, or emotional word valence that can change according to context and domain, although several works adopt this simplified methodology which in practice can work adequately[55].

In conclusion, ML predictivity shows a specific emotion pattern strictly related to a specific medical condition when "successful" medical web pages are concerned; in the case of the study, disgust was shown to be the single basic emotion with more relevant discriminative power between the spine and great joint degenerative conditions, outlining a different psychosocial profile. Different explanatory hypotheses have been proposed, to be investigated in future research. The notion of DACC is outlined as a conceptual framework. Future research should consider other languages for comparison, sentiment analysis of digital information produced by patients, i.e., through social networks, or confrontation of digital emotional content to real-life emotional frames; they may also focus on other sectors of DACC and pathologies, possibly leading to new behavioral analysis models. Exploiting such knowledge holds the potential to overcome limitations of treatments based on mechanistic pathogenetic reductionism, enforcing a modern BPS model that includes DACC role; the process may lead to a more comprehensive understanding of how internet information spreads, drive application of modern neuroscience emerging evidence to medical practice, optimizing approaches to major medicinal issues, pursue correct information to the public, ultimately leading to optimization of cures and greater prosperity of the community.

Material and methods

Identifying relevant sources. Starting point keywords were identified by the three authors who specialized in the spine (DC), knee (EK), or hip surgery (RF), as outlined in the first block in Fig. 7. Relevant internet sources were analyzed using SEMrush Competitors Research (SEMrush), a software designed for companies to run digital marketing. SEMrush can identify trends that occur within a web niche and rank performance on a content-specific base. The following parameters were evaluated for each internet source, as outlined in the second block for Fig. 7.

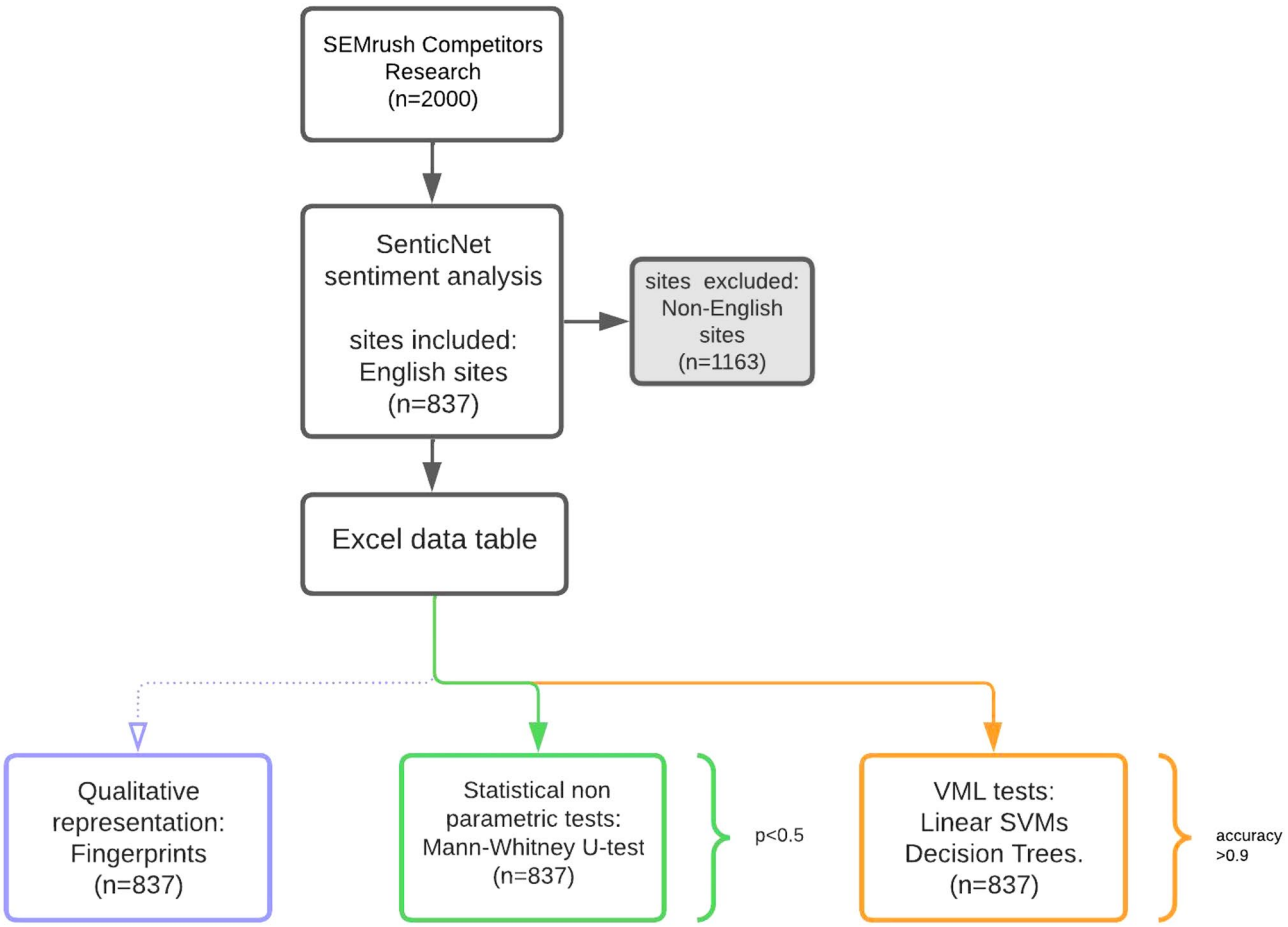


Figure 7. Block diagram, identifying the steps of the workflow. The number of sites to be included in the initial analysis was arbitrarily set in SEMrush to 2000. Senticnet extracted the count for each pool of words classified as pertaining to one emotion. We plotted the relative occurrence of pooled emotion-related words (columns) with each site (rows) for each condition (one spreadsheet for condition) in Excel, for English sites. The fingerprint graphic format was developed for a qualitative representation of each condition emotional pattern. The variables were confronted on a statistical level (Mann–Whitney U-test) for association significance and were forwarded to SVM linear testing.

(1) Page AS: SEMrush standard metric used to measure the overall quality of the URL and influence on SEO. The score is based on the number of backlinks, referring domains, organic search traffic, and other cases. It's a tool to measure the impact of a webpage or domain links. Authority Score is a compound domain score that grades the overall quality of a website. The higher the score, the more assumed weight a domain or webpage backlinks could have.
(2) Ref Domain: is a website that links out to another website whose backlink profile you analyze. When Google measures a domain's trust from its backlinks, the search engine weighs having a high number of referring domains. Note: the total number of referring domains that have at least one link pointing to a given URL. SEMrush only considers the domains it has seen in the last few months.
(3) Backlinks are links from one website to another. Search engines like Google use backlinks as a ranking signal. Note: total number of backlinks pointing to a given URL. SEMrush only considers the backlinks it has seen in the last few months. For clarity, for a given web resource, a Backlink is a link from some other website (the referrer) to that web resource (the referent). A web resource may be (for example) a website, web page, or web directory. A backlink is a reference comparable to a citation.
(4) Search Traffic: The term "search traffic" refers to the entire traffic from various visitor sources through a specific medium. Note: the amount of estimated organic traffic brought to a given URL with the keyword analyzed for a given time interval.
(5) URL Keyword: The number of keywords for which a given URL ranks in search results.

Semantic analysis—Senticnet. Content and data from URLs were converted to raw text and exported as a CSV file. The gathered data contains categories of orthopedics diseases or health conditions (back pain, hip prosthesis, knee prosthesis, etc.). Sentiment analysis (opinion mining, emotion AI) was used to extract words related to emotions. In the present study used SenticNet (https://sentic.net) a multi-disciplinary approach to opinion mining at the crossroads between affective and common-sense computing that combines semiotics, psychology, linguistics, and machine learning elements; the step is outlined as the third block in Fig. 7. Sentic computing, as opposed to statistical sentiment analysis, is a multi-disciplinary paradigm that focuses on a semantic-preserving representation of natural language concepts and sentence structure. SenticNet is based on the Hourglass of Emotions, an emotion categorization model developed to properly express the affective information associated with natural language text[56]. Using this categorization, feelings are reorganized around four independent dimensions with different levels of activation that make up the total emotional state of mind. Affective states are classified into four dimensions—Pleasantness, Attention, Sensitivity, and Attitude. Each of the four affective dimensions is characterized by six levels of activation, called "sentic levels' ", which determine the intensity of the emotion. In the *Aptitude* dimension can be found l*oathing, disgust, boredom, acceptance, trust, admiration*. In *Pleasantness*, *grief, sadness, pensiveness, serenity, joy, ecstasy*. In *Sensitivity* dimension, *terror, fear, apprehension, annoyance, anger, and rage*. Finally, in the *Attention* state, there are *amazement, surprise, distraction, interest, anticipation, and vigilance*.

BabelSenticNet[57] is a multilingual concept-level knowledge base for sentiment analysis based on SenticNet for emotion recognition and the output returned us joy, admiration, surprise, fear, disgust, anger, sadness, and interest.

For sentiment analysis, Natural language processing (NLP) was used.

The system can then extract accurate information and insights from the papers and categorize and organize them[58]. The text has been prepared with three processes.

- Tokenization is the process of breaking down a given text into the smallest element in a sentence, termed a token. Output:" Hip"," replacement"," surgery", "can"," help", "relieve"
- Lemmatization, the process of discovering the normal form of an original word in the dictionary
- Part of Speech Tagging, Labeling words in a text according to their word kinds is known as Part of Speech Tagging (POS-Tag, that is noun, adjective, adverb, verb, etc.). It is a method for transforming a sentence into a list of words or tuples

Then with Sentiment analysis, a systematic identification, extraction, quantification, and study of affective states were achieved. In addition to the emotion variables, other quantities such as the number of sentences, words, content words, etc. were stored in a spreadsheet.

The documents were grouped by condition. Each partial dataset of documents from a particular group was modeled by a matrix obtained by stacking the emotional vectors. Columns pertained to emotions and rows indexed documents.

Nonparametric statistical tests and machine learning. Comparisons were performed in the English language documents with statistical nonparametric tests (Mann–Whitney U-test, repeated three times) and by verifying health condition predictivity using different ML binary classifiers: Naive Bayes, Multi-Layer Perceptrons (MLP), Support Vector Machines (SVM) with several kernels, Decision Trees, and eXtreme Gradient Boosting (XGBoost). The classification quality results were comparable, so we concentrated on Linear SVMs, which offer optimal linear feature-space partitioning[59], and Decision Trees, because they provide explainable results[60]. Using ML to classify the documents allowed us to find many cases with classification accuracy (assessed with a fivefold cross-validation scheme) higher than 0.90, which indicates that the document's emotional contents have peculiar patterns, in each class of a pair of health conditions, making them different and recognizable.

For evaluating performance, we computed various statistics from the confusion matrices, in particular *sensitivity (recall), specificity, precision,* and *accuracy* metrics. To also find out the optimal blend of recall and precision,

we combined these two metrics in the F1 score, that is the harmonic mean of precision and recall taking both metrics into account as follows:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

The harmonic mean is used instead of a simple average because the former punishes extreme values (a classifier with a precision of 1.0 and a recall of 0.0 has a simple average of 0.5 but an F1 score of 0).

This step is referenced in the last block of the diagram in Fig. 7.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding authors upon reasonable request.

Code availability

The code used to generate results reported in the manuscript that are central to the main claims is available from the corresponding authors upon reasonable request.

Received: 11 September 2022; Accepted: 16 March 2023
Published online: 21 March 2023

References

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**(10159), 1789–1858 (2018).
2. Eisenstein, S. M., Balain, B. & Roberts, S. Current treatment options for intervertebral disc pathologies. *Cartilage.* **11**(2), 143–151 (2020).
3. Deyo, R. A., Von Korff, M. & Duhrkoop, D. Opioids for low back pain. *BMJ* **350**, g6380 (2015).
4. Bonnie, R. J., Schumacher, M. A., Clark, J. D. & Kesselheim, A. S. Pain management and opioid regulation: Continuing public health challenges. *Am. J. Public Health.* **109**(1), 31–34 (2000).
5. Phillips, K. & Clauw, D. J. Central pain mechanisms in chronic pain states—Maybe it is all in their head. *Best Pract. Res. Clin. Rheumatol.* **25**(2), 141–154 (2011).
6. Buchbinder, R., Richards, B. & Harris, I. Knee osteoarthritis and role for surgical intervention: Lessons learned from randomized clinical trials and population-based cohorts. *Curr. Opin. Rheumatol.* **26**, 138–144 (2014).
7. Maher, C., Underwood, M. & Buchbinder, R. Non-specific low back pain. *Lancet* **389**(10070), 736–747 (2017).
8. Thomson, S. Failed back surgery syndrome—Definition, epidemiology and demographics. *Br. J. Pain* **7**(1), 56–59 (2013).
9. Blyth, F. M., Briggs, A. M., Schneider, C. H., Hoy, D. G. & March, L. M. The global burden of musculoskeletal pain—Where to from here?. *Am. J. Public Health* **109**(1), 35–40 (2019).
10. Le Borgne, M., Boudoukha, A. H., Petit, A. & Roquelaure, Y. Chronic low back pain and the transdiagnostic process: How do cognitive and emotional dysregulations contribute to the intensity of risk factors and pain?. *Scand. J. Pain* **17**, 309–315 (2017).
11. Wertli, M. M., Rasmussen-Barr, E., Weiser, S., Bachmann, L. M. & Brunner, F. The role of fear avoidance beliefs as a prognostic factor for outcome in patients with nonspecific low back pain: A systematic review. *Spine J.* **14**(5), 816–36.e4 (2014).
12. Alyousef, B. *et al.* Negative beliefs about back pain are associated with persistent, high levels of low back disability in community-based women. *Menopause (New York, N.Y.)* **25**(9), 977–984 (2018).
13. Burns, J. W., Bruehl, S. & Quartana, P. J. Anger management style and hostility among patients with chronic pain: Effects on symptom-specific physiological reactivity during anger- and sadness-recall interviews. *Psychosom. Med.* **68**(5), 786–793 (2006).
14. Unver, B., Ertekin, Ö. & Karatosun, V. Pain, fear of falling and stair climbing ability in patients with knee osteoarthritis before and after knee replacement: 6 month follow-up study. *J. Back Musculoskelet. Rehabil.* **27**(1), 77–84 (2014).
15. Moore, A., Eccleston, C. & Gooberman-Hill, R. "It's Not My Knee": Understanding ongoing pain and discomfort after total knee replacement through re-embodiment. *Arthritis Care Res.* **74**(6), 975–981 (2022).
16. Pan, F., Tian, J., Aitken, D., Cicuttini, F. & Jones, G. Predictors of pain severity trajectory in older adults: A 10.7-year follow-up study. *Osteoarthritis Cartilage* **26**(12), 1619–1626 (2018).
17. Nwankwo, V. C. *et al.* Resilience and pain catastrophizing among patients with total knee arthroplasty: A cohort study to examine psychological constructs as predictors of post-operative outcomes. *Health Qual. Life Outcomes* **19**(1), 136 (2021).
18. Sachs, M. E., Damasio, A. & Habibi, A. The pleasures of sad music: A systematic review. *Front. Hum. Neurosci.* **9**, 404 (2015).
19. Brown, C.A. Pain and complex adaptive system theory. in *Handbook of Systems and Complexity in Health*. 397–421 SpringerLink, (2013).
20. Sander, D. Models of emotion: The affective neuroscience approach. In *The Cambridge Handbook of Human Affective Neuroscience* (eds Armony, J. & Vuilleumier, P.) 5–53 (Cambridge University Press, 2013).
21. Hilbert, M. & López, P. The world's technological capacity to store, communicate, and compute information. *Science* **332**(6025), 60–65 (2011).
22. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), p107-117 (1998).
23. Hamborg, F., Donnay, K. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (2021).
24. Dolce, P., Marocco, D., Maldonato, M. N. & Sperandeo, R. Toward a machine learning predictive-oriented approach to complement explanatory modelling. An application for evaluating psychopathological traits based on affective neurosciences and phenomenology. *Front. Psychol.* **11**, 446 (2020).
25. Engel, G. L. The need for a new medical model: A challenge for biomedicine. *Science* **196**(4286), 129–136 (1977).
26. Damasio, A. R., Tranel, D. & Damasio, H. C. Somatic markers and the guidance of behaviour: Theory and preliminary testing. In *Frontal Lobe Function and Dysfunction* (eds Levin, H. S. *et al.*) 217–229 (Oxford University Press, 1991).
27. Damasio, A. The feeling of what happens: body and emotion in the making of consciousness harvest (2000).
28. Cauda, F., Costa, T., Diano, M., Duca, S. & Torta, D. M. Beyond the "Pain Matrix," inter-run synchronization during mechanical nociceptive stimulation. *Front. Hum. Neurosci.* **8**, 265 (2014).
29. Doleys, D. M. Chronic pain as a hypothetical construct: A practical and philosophical consideration. *Front. Psychol.* **8**, 664 (2017).
30. Sturmberg, J. P. Health and Disease Are Dynamic Complex-Adaptive States Implications For Practice And Research. *Front. Psych.* **12**, 595124 (2021).

31. Trinderup, J. S., Fisker, A., Juhl, C. B. & Petersen, T. Fear avoidance beliefs as a predictor for long-term sick leave, disability and pain in patients with chronic low back pain. *BMC Musculoskelet. Disord.* **19**(1), 431 (2018).
32. Bruehl, S., Burns, J. W., Chung, O. Y., Ward, P. & Johnson, B. Anger and pain sensitivity in chronic low back pain patients and pain-free controls: The role of endogenous opioids. *Pain* **99**(1–2), 223–233. https://doi.org/10.1016/s0304-3959(02)00104-5 (2002).
33. Maradit Kremers, H. *et al.* Prevalence of total hip and knee replacement in the United States. *J. Bone Joint Surg. Am.* **97**(17), 1386–1397. https://doi.org/10.2106/JBJS.N.01141 (2015).
34. Wylde, V. *et al.* Preoperative widespread pain sensitization and chronic pain after hip and knee replacement: A cohort analysis. *Pain* **156**(1), 47–54 (2015).
35. Vlaeyen, J. & Linton, S. J. Fear-avoidance and its consequences in chronic musculoskeletal pain: A state of the art. *Pain* **85**(3), 317–332 (2000).
36. Oaten, M. J., Stevenson, R. J. & Case, T. I. The effect of disgust on pain sensitivity. *Physiol. Behav.* **138**, 107–112 (2015).
37. Loehle, C. Social barriers to pathogen transmission in wild animal populations. *Ecology* **76**, 326–335 (1995).
38. Jones, A. & Fitness, J. Moral hypervigilance: The influence of disgust sensitivity in the moral domain. *Emotion* **8**(5), 613–627 (2008).
39. Horberg, E. J., Oveis, C., Keltner, D. & Cohen, A. B. Disgust and the moralization of purity. *J. Pers. Soc. Psychol.* **97**(6), 963–976 (2009).
40. Sherman, G. D. & Haidt, J. Cuteness and disgust: The humanizing and dehumanizing effects of emotion. *Emot. Rev.* **3**(3), 245–251 (2011).
41. Beadle, J. N. & De la Vega, C. E. Impact of aging on empathy: Review of psychological and neural mechanisms. *Front. Psych.* **10**, 331 (2019).
42. Jordan, S. E. *et al.* Using twitter for public health surveillance from monitoring and prediction to public response. *Data.* **4**(1), 6. https://doi.org/10.3390/data4010006 (2019).
43. Shah, A. M., Naqvi, R. A. & Jeong, O. R. Detecting topic and sentiment trends in physician rating websites: Analysis of online reviews using 3-wave datasets. *Int. J. Environ. Res. Public Health* **18**(9), 4743 (2021).
44. DataReportal (2022), "Digital 2022 Global Digital Overview". https://datareportal.com/reports/digital-2022-global-overview-report (2022). (Accessed 14 July 2022)
45. Wald, H. S., Dube, C. E. & Anthony, D. C. Untangling the web—The impact of Internet use on health care and the physician–patient relationship. *Patient Educ. Couns.* **68**(3), 218–224 (2007).
46. Bzdok, D., Varoquaux, G. & Steyerberg, E. W. Prediction, not association, paves the road to precision medicine. *JAMA Psychiat.* **78**(2), 127–128 (2021).
47. Goldstein, P. *et al.* Emerging clinical technology: Application of machine learning to chronic pain assessments based on emotional body maps. *Neurotherapeutics* **17**(3), 774–783 (2020).
48. Combs, A. & Krippner, S. Collective consciousness and the social brain. *J. Conscious. Stud.* **15**(10–11), 264–276 (2008).
49. Goldenberg, A., Garcia, D., Halperin, E. & Gross, J. J. Collective emotions. *Curr. Dir. Psychol. Sci.* **29**(2), 154–160 (2020).
50. Cheok, A. D. *Hyperconnectivity and the Future of Internet Communication* (Lambert Academic Publishing, 2015).
51. Boire, R. G. On cognitive liberty (part I). *J. Cognit. Liberties.* **1**(1), 7–13 (2000).
52. Goldenberg, A. & Gross, J. J. Digital emotion contagion. *Trends Cogn. Sci.* **24**(4), 316–328 (2020).
53. Ahmadvand, A., Gatchel, R., Brownstein, J., Nissen, L. The biopsychosocial-digital approach to health and disease: Call for a paradigm expansion. *J. Med. Internet Res.* **20**(5) (2018).
54. Smith, J. A. *et al.* Do people with low back pain walk differently? A systematic review and meta-analysis. *J Sport Health Sci.* **11**(4), 450–465 (2022).
55. Samothrakis, S. & Fasli, M. Emotional sentence annotation helps predict fiction genre. *PLoS ONE* **10**(11), e0141922 (2015).
56. Cambria, E. & Hussain, A. *Sentic Computing: Techniques, Tools, and Applications* (Springer, 2012).
57. Vilares Peng, D.H., Satapathy, R., Cambria, E. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1292–1298 (2018).
58. Feldman, S.E. NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. Online, 23 (1999).
59. Suthaharan, S. Support vector machine. in *Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems*, Vol. 36 (2016).
60. Islam, M. R., Ahmed, M. U., Barua, S. & Begum, S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl. Sci.* **12**(3), 1353 (2022).

Author contributions

D.C.—original idea, design of the study, interpretation of results, drafting and editing the final manuscript. S.B.—interpretation of results, drafting, and editing. L.C.—machine learning and statistical methods. M.S.A., L.A., V.B., M.M.H., A.M.—sentiment analysis. P.H.—critical revision. R.F., E.K.—supervision of the work, critical revision. G.N.—machine learning methods, final revision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.C. or S.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

B.2 Published Article 2: Exploring sentiments in summarization: SentiTextRank, an Emotional Variant of TextRank.

This study, authored by Md Murad Hossain, Luca Anselma, and Alessandro Mazzei, and published in CEUR WORKSHOP PROCEEDINGS, introduces SentiTextRank, an innovative approach to integrating sentiment analysis into the TextRank algorithm for text summarization.

The research explores the role of sentiment in enhancing the relevance and emotional tone of summaries, particularly within the field of natural language processing and information retrieval. By proposing SentiTextRank, we aim to advance the effectiveness of automated summarization systems, thereby enriching the comprehension and usability of textual content across various domains.

Exploring sentiments in summarization: SentiTextRank, an Emotional Variant of TextRank

Conference Paper · December 2023

CITATIONS
0

READS
15

3 authors:

Md. Murad Hossain
Bangabandhu Sheikh Mujibur Rahman Science & Technology University
38 PUBLICATIONS **97** CITATIONS
SEE PROFILE

Luca Anselma
Università degli Studi di Torino
79 PUBLICATIONS **534** CITATIONS
SEE PROFILE

Alessandro Mazzei
Università degli Studi di Torino
97 PUBLICATIONS **503** CITATIONS
SEE PROFILE

Exploring sentiments in summarization: SentiTextRank, an Emotional Variant of TextRank

Md. Murad Hossain[1,*,†], Luca Anselma[1,†] and Alessandro Mazzei[1,†]

[1] *Department of Computer Science, University of Turin, Corso Svizzera 185, 10149 Torino, Italy*

Abstract

English. A summary that aims at preserving the emotions of the original text can be interesting in certain application scenarios, such as in the generation of metareviews, both in academic and commercial domains. TextRank is a well-studied algorithm for automatic extractive summarization. This work introduces SentiTextRank, an emotional variant of TextRank, to enhance the extractive technique for both single-document and multi-document summarization. SentiTextRank incorporates emotions into the summarization process by classifying sentences into the eight emotional categories used in SenticNet. The preliminary evaluation of SentiTextRank yields encouraging results. In particular, our method generates informative summaries composed of sentences that preserve the emotional content of the original document.

Italian. Un riassunto che miri a preservare le emozioni del testo originale può essere interessante in alcuni scenari applicativi, come ad esempio nella generazione di meta-recensioni sian nel dominio accademico che in quello commerciale. TextRank è un algoritmo per il riassunto automatico estrattivo molto studiato. Questo lavoro introduce SentiTextRank, una variante emozionale di TextRank, per potenziare la tecnica estrattiva sia per il riassunto di singoli documenti che per il riassunto di documenti multipli: SentiTextRank integra le emozioni nel processo di sintesi, classificando le frasi nelle otto categorie emotive utilizzate in SenticNet. La valutazione preliminare di SentiTextRank produce dei risultati incoraggianti. In particolare, il nostro metodo produce dei riassunti informativi formati da frasi che rispettano il contenuto emozionale del documento originale.

Keywords

Extractive summarization, SentiTextRank, emotional variant, single and multi-document Summary, emotional content.

1. Introduction

Summarization is the process of reducing a larger body of information into a concise and coherent summary that captures the essential points and main ideas. Extractive summarization involves selecting and combining sentences or phrases directly from the source text to form the summary [1] and plays an important role in condensing news articles into concise summaries, allowing readers to quickly grasp the key information. Traditional extractive methods primarily rely on lexical word distance to select important sentences for summarization. In many cases the emotional aspects found in the documents are not considered in summarization, and this can affect how readers engage with and understand the information.

Sentiment analysis is the extraction of subjective information from text, encompassing emotions and opinions, and the classification based on the expressed emotions, such as happiness, sadness, anger, fear, or surprise, to capture the overall emotional sentiment [2] and [3]. Sentiment analysis had a huge impact on many applications of NLP in the last years, but there is still space for understanding the details of its implementations [4].

The current study employs SenticNet [3], a multidisciplinary approach to opinion mining that lies at the intersection of affective and common-sense computing. This approach integrates elements from semiotics, psychology, linguistics, and machine learning. Unlike statistical sentiment analysis, Sentic computing focuses on preserving the semantic representation of natural language concepts and sentence structure. The foundation of SenticNet is the Hourglass of Emotions, an emotion categorization model designed to accurately express the affective information present in natural language text.

To the best of our knowledge, sentiment generation is an understudied argument in the field of automatic summarization. Despite the advancements in text summarization techniques, there is a gap in research when it comes to considering emotions in the process. However, we believe that there are a number of applications in which the emotions in a summary should correspond to the emotions of the original document(s). For instance, this is the case of meta-reviews in conference management or the summarization of product reviews in the case of e-commerce. For these application domains, it is

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

* Corresponding author.

† These authors contributed equally.

mdmurad.hossain@unito.it (Md. M. Hossain); luca.anselma@unito.it (L. Anselma); alessandro.mazzei@unito.it (A. Mazzei)

https://shorturl.at/bEGZ0 (Md. M. Hossain); http://www.di.unito.it/~anselma/ (L. Anselma); http://www.di.unito.it/~mazzei/ (A. Mazzei)

0000-0002-8224-3246 (Md. M. Hossain); 0000-0003-2292-6480 (L. Anselma); 0000-0003-3072-0108 (A. Mazzei)

important to consider emotions to create summaries that truly reflect the essence of the source texts.

In this paper, by incorporating sentiment scoring between sentences, we generate summaries that capture the emotional tone and impact of the original text. We believe that exploring this aspect further would lead to more comprehensive and effective text summarization methods.

This paper has two main goals. First, we define a new algorithm called *SentiTextRank*, which is an *emotional* variant of TextRank [5]. Second, we provide an initial evaluation of SentiTextRank by considering two automatic metrics based on content distance.

Note that modern LLMs showed some abilities in summarizing texts by using a specific style, with some limitations in producing a summary that is truly extractive. Moreover, LLMs showed also a big impact from the point of view of the required computational resources. We believe that the work presented in this paper, that requires just few hours to conduct all the experiments, can be seen as a cheap (in many senses) alternative to the use of modern expensive (in many senses) LLMs[1].

The paper is structured as follows. In Section 2, we define the new SentiTextRank algorithm, in Section 3 we report the result of a first experimental evaluation of the SentiTextRank algorithm and the Section 4 ends the paper pointing out to work in progress.

2. SentiTextRank: a variant of TextRank accounting for emotions

TextRank is a popular algorithm for extractive summarization which constructs a graph of sentences or words from a text and assigns scores to each node based on their importance in the graph structure. Finally, it ranks the nodes and selects the top-ranked sentences or words as the summary [5]. The TextRank algorithm is based on the PageRank algorithm, where the sentences of the documents play the role of web pages, and a similarity score plays the role of hyperlink connectivity. Our approach enhances traditional TextRank by incorporating emotions. In particular, we categorize sentences of the original source(s) on the basis of emotions using SenticNet [6]. On the basis of this classification, we obtain a number of distinct emotion sets of sentences. The main idea is to build one single final summary by merging in a selective way the results of TextRank on each one of these emotions sets.

So, the proposed *SentiTextRank* algorithm generates extractive summaries with emphasis on emotion categories through the following steps:

[1]We thank an anonymous reviewer for pointing out this point.

SentiTextRank: Input=Source, **Output**=Sum_F

1. Set compression ratio parameter C between the source(s) and the final summary Sum_F.
2. Classify sentences of the source(s) into different SenticNet emotion categories CAT_{em} with $em \in$ {joy, admiration, surprise, fear, disgust, anger, sadness, interest}.
3. Generate a summary Sum_{em} for each emotion category CAT_{em} by using TextRank.
4. Build Sum_F by picking a number of sentences proportional to C from each Sum_{em} maintaining the original sentence order of the source document.

3. Experimental Result and Discussion

In this section, we present the experimental results of single-document summarization using two datasets, the CNN/Daily Mail dataset (CNN) and the DUC2001 single document dataset (D01), as well as the results of multi-document summarization using two datasets, the DUC2001 multi-document dataset (MD01) and the DUC2004 multi-document dataset (MD04).

The DUC 2001 single document and DUC 2001 Multi-document datasets were collected from the website[2] and consist of news datasets. For our experiments with single documents, we utilized a sample data set of 54 documents. The DUC 2004 multi-document summarization dataset[3] includes 50 items with multiple files and four reference files per item, from which we utilized the first reference for each item. Additionally, we used the CNN/Daily mail dataset[4], where we considered the "highlights" column as the reference summary. Our experiments were conducted on the first 100 rows of text in the CNN/Daily mail dataset. Since the datasets provide just abstractive gold summaries, in order to provide a fair comparison we have converted the abstractive summaries into extractive summaries. This procedure has been proposed in [7, 8]. An extractive reference summary should yield the highest Rouge score when compared to the gold abstractive summary. As finding the globally optimal subset of sentences that maximize the Rouge score is computationally intractable, we adopt a greedy approach: we iteratively add one sentence at a time to the summary, ensuring that the Rouge score of the current set of selected sentences is maximized in relation to the entire gold summary. We repeat this process until there are no more candidate sentences that could enhance the Rouge score when added

[2]https://duc.nist.gov/data.html
[3]https://rb.gy/gp1gbt
[4]https://rb.gy/v4u2g

Original Text	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk.
Gold Abstractive Summary	Experts question if packed out planes are putting passengers at risk. U.S consumer advisory group says minimum space must be stipulated.
Reference Extractive Summary	Ever noticed how plane seats appear to be getting smaller and smaller?. This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans.
Lead	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putting passengers at risk.
TR	They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. This week, a U.S consumer advisory group set up by the Department of Transportation said at a public hearing that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans.
STR	They say that the shrinking space on aeroplanes is not only uncomfortable - it's putting our health and safety in danger. 'It is time that the DOT and FAA take a stand for humane treatment of passengers.

Table 1
An excerpt from the Original Text from the CNN dataset, the existing reference summary (Gold Abstractive Summary), the generated reference summary (Reference Extractive Summary), the lead baseline (Lead), the summary generated by TextRank (TR), and the summary generated by SentiTextRank (STR).

to the current summary set. The subset of sentences that we have at this point is then considered the extractive reference summary for the evaluation. In Table 1 we report an example of summaries generated with the different methods.

Table 1 reports an excerpt from the CNN dataset (Original Text) and the corresponding reference summary (Gold Abstractive Summary). Moreover, Table 1 contains the corresponding generated reference extractive summary (Reference Extractive Summary), the prefix baseline (Lead), the text generated with the TextRank baseline (TR) and, finally, the SentiTextRank generated summary (STR).

Dataset	Algorithm	RL-F1	BERT-F1
D01	Lead	**0.600**	**0.729**
	TR	*0.382*	*0.649*
	STR	0.366	0.605
CNN	Lead	**0.711**	**0.794**
	TR	0.345	*0.642*
	STR	*0.372*	0.608
MD01	Lead	**0.802**	**0.851**
	TR	0.061	*0.560*
	STR	*0.163*	0.505
MD04	Lead	**0.511**	**0.683**
	TR	0.123	*0.575*
	STR	*0.227*	0.542

Table 2
The results of summarization experiments. Lead = Lead Baseline, TR = TextRank, STR = SentiTextRank.

Table 2 presents the experimental results of different summarization methods, namely the Baseline (Lead), TextRank (TR), and our proposed method, SentiTextRank (STR) evaluated on single-document datasets DUC-2001 (D01) and CNN, and the multi-document datasets DUC-2001 (MD01) and DUC-2004 (MD04). As a baseline, we selected the leading sentences from the original documents based on the compression ratio. We evaluated the summaries using two measures: Rouge-L F1 (RL-F1) and BERT F1. ROUGE (Recall Oriented Understudy for Gisting Evaluation) is frequently used to assess how well summarization techniques perform. Rouge-L computes ROUGE for the longest sequence of n-grams [9]. BERT F1 score is a metric commonly used in text classification tasks. It measures the token-level similarity between the generated summary and the reference summary, considering both precision and recall [10].

The results consistently indicate that the Lead method outperforms the other methods across the datasets, showcasing its superiority in generating high-quality summaries. Specifically, Lead achieves the highest scores in Rouge-L F1 and BERT-F1 for D01, CNN, MD01, and MD04. The TR and STR methods exhibit moderate performance in specific evaluation metrics.

Note that the better performance of the Lead method can be attributed to the fact that all the experiments were conducted using news datasets; indeed this result is consistent with the results reported in the literature, where a baseline composed of the leading sentences frequently outperforms extractive and abstractive models on news datasets [11]. However, we think that the comparison between original TR and STR shows encouraging results.

Indeed, the fact that using emotions does not degrade

the performance with regard to TextRank shows that we can produce a summary that represents the content as well as the emotions of the source documents. In order to experimentally prove this intuition, we need to formalize an *emotional distance* between summaries and source documents. We plan to develop this point in the future by both (1) using LLM and (2) considering human evaluation.

Further research is necessary to evaluate the performance of our proposed STR method on another domain dataset to provide a comprehensive understanding of its effectiveness.

4. Conclusion and Future Work

This paper introduces the SentiTextRank algorithm, which integrates emotions into the extractive summarization process to create more informative and emotionally rich summaries. The experimental results are encouraging with respect the effectiveness of SentiTextRank in capturing factual information.

The ongoing work on SentiTextRank is following different directions.

First, we want to design a new version of the algorithm that will not be based on the classification of a sentence in one single prevalent emotion. The idea that we want to develop is to define one single measure that combines both *content* and *emotion* similarities. By using this combined measure, we can apply the original TextRank algorithm on the entire set of sentences from the source(s) and obtain one single ranking structure accounting for both content and emotion.

Second, we want to conduct more extensive experiments also on datasets from different domains. In particular, we are considering medical applications since the affective component of medical information can represent a relevant biopsychosocial feature [12].

Third, we are aware that automatic metrics not always measure a real *quality* of the summarized text with respect to human judgment [13, 14]. So, we plan in future to conduct human-based evaluation too.

References

[1] A. Nenkova, K. McKeown, et al., Automatic summarization, Foundations and Trends® in Information Retrieval 5 (2011) 103–233.

[2] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artificial Intelligence Review 55 (2022) 5731–5780.

[3] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective computing and sentiment analysis, A practical guide to sentiment analysis (2017) 1–10.

[4] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai, et al., Sentiment analysis: It's complicated!, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1886–1895.

[5] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.

[6] E. Cambria, Y. Li, F. Z. Xing, S. Poria, K. Kwok, Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis, in: Proceedings of the 29th ACM international conference on information & knowledge management, 2020, pp. 105–114.

[7] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 3075–3081.

[8] M. Isonuma, T. Fujino, J. Mori, Y. Matsuo, I. Sakata, Extractive summarization using multi-task learning with document classification, in: Proceedings of the 2017 Conference on empirical methods in natural language processing, 2017, pp. 2101–2110.

[9] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[11] Y. Liu, M. Lapata, Text summarization with pretrained encoders, arXiv preprint arXiv:1908.08345 (2019).

[12] D. Caldo, S. Bologna, L. Conte, M. S. Amin, L. Anselma, V. Basile, M. M. Hossain, A. Mazzei, P. Heritier, R. Ferracini, et al., Machine learning algorithms distinguish discrete digital emotional fingerprints for web pages related to back pain, Scientific Reports 13 (2023) 4654.

[13] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why we need new evaluation metrics for nlg, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2241–2252. doi:10.18653/v1/D17-1238.

[14] F. Moramarco, A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, A. Savkov, Human evaluation and correlation with automatic metrics in consultation note generation, in: Proceedings of the 60th Annual Meeting of the Asso-

ciation for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5739–5754. URL: https://aclanthology.org/2022.acl-long.394. doi:10.18653/v1/2022.acl-long.394.

B.3 Published Article 3: Anticipating User Intentions in Customer Care Dialogue Systems.

This research, published in the IEEE Transactions on Human-Machine Systems, explores advanced methodologies for anticipating user intentions within customer care dialogue systems. Authored by a team of researchers, including Alessandro Mazzei, Luca Anselma, Manuela Sanguinetti, Amon Rapp, Dario Mana, Md Murad Hossain, Viviana Patti, Rossana Simeoni, and Lucia Longo, the study delves into the intricate dynamics of human-machine interactions, to enhance the responsiveness and effectiveness of automated customer service. By employing sophisticated computational techniques, the paper addresses critical challenges in understanding and predicting user intents, contributing pivotal insights to the field of human-machine systems and artificial intelligence applications in customer service contexts.

Anticipating User Intentions in Customer Care Dialogue Systems

Alessandro Mazzei, Luca Anselma, Manuela Sanguinetti, Amon Rapp, Dario Mana, Md. Murad Hossain, Viviana Patti, Rossana Simeoni, and Lucia Longo

***Abstract*—In this article, we investigate the case of human-machine dialogues in the specific domain of commercial customer care. We built a corpus of conversations between users and a customer-care chatbot of an Italian Telecom Company, focusing on a sample of conversations where users contact the service asking for explanations about billing issues or overcharges. We observed that users' requests are often vague, generic or incomprehensible. In such cases, commercial dialogue systems typically ask for clarifications or further details to fully understand users' specific requests. However, from the corpus analysis it appeared that chatbot's clarifying requests may result in ineffective interactions, with users eventually giving up the conversation or switching to a human agent for a faster query resolution. A recovery strategy is thus needed to anticipate users' information needs, or intentions. We address this issue resorting to GEN-DS, a dialogue system based on symbolic data-to-text generation. GEN-DS analyzes the user-company contextual relational knowledge, with the aim to generate more relevant answers to unclear questions. In this article, we describe the GEN-DS architecture along with the experiments we carried out to evaluate its output. Results from an offline human evaluation show significant improvements of GEN-DS compared to the original system. These improvements concern properties such as utility, necessity, understandability, and quickness of the information communicated in the dialogue. We believe that GEN-DS techniques may find application in all the dialogue systems that need to manage vague requests and must rely on relational knowledge.**

***Index Terms*—Human-computer interface, man-machine systems, natural language processing.**

Manuscript received 20 May 2021; revised 22 January 2022 and 8 April 2022; accepted 28 May 2022. Date of publication 15 August 2022; date of current version 15 September 2022. This work was supported in part by the TIM s.p.a. Studi e Ricerche su Sistemi Conversazionali Intelligenti under Grant CENF_CT_RIC_19_01. This article was recommended by Associate Editor M. Hou. *(Corresponding author: Alessandro Mazzei.)*

Alessandro Mazzei, Luca Anselma, Amon Rapp, Md. Murad Hossain, and Viviana Patti are with the Dipartimento di Informatica, Università degli Studi di Torino, 10124 Torino, Italy (e-mail: alessandro.mazzei@unito.it; anselma@di.unito.it; amon.rapp@unito.it; md.hossain50@edu.unito.it; viviana.patti@unito.it).

Manuela Sanguinetti is with the Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, 09124 Cagliari, Italy (e-mail: manuela.sanguinetti@unica.it).

Dario Mana, Rossana Simeoni, and Lucia Longo are with the Telecom Italia Mobile, 10156 Torino, Italy (e-mail: dario.mana@telecomitalia.it; rossana.simeoni@telecomitalia.it; lucia.longo@telecomitalia.it).

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/THMS.2022.3184400.

Digital Object Identifier 10.1109/THMS.2022.3184400

I. Introduction

DIALOGUE systems (DSs) may have different forms and pursue different aims. Grudin and Jacques [1] proposed a taxonomy of DSs based on their conversational focus: *virtual companions* usually engage on any topic and keep the conversation going, while *intelligent assistants* converse on any topic but aim to keep the conversation short; instead, *task-oriented agents* aim to solve specific problems and are addressed to perform short conversations.

An ever-increasing number of companies are adopting task oriented DSs as a preferred way of interacting with their customers. DSs provide several benefits to companies, to their customers, and to the customer care human operators (e.g., [2]–[4]). They are available 24/7 and can keep the context of an ongoing conversation for hours or even days, allowing users to solve their problems even when they get distracted from the support session. Moreover, DSs allow the collection, directly from customers, of unconstrained natural language text, which may then be interpreted through computational linguistics techniques [5]. They would thus provide an extremely valuable source of knowledge about customers' expectations, preferences, and behavior.

Customers commonly approach the DS with a variety of attitudes and expectations (e.g., [6], [7]), which are often not met by the technology. Users report a satisfying experience when the agent can correctly interpret their requests, provide appropriate and relevant responses, and communicate clearly what it can do [8]. Users holding high expectations toward the DS often approach it as if it were a human operator and explain their situation and problem at length, including details that give a lot of contextual information, but may not be directly useful (or usable) by an automatic system. It appears that these kinds of users get easily frustrated or angry, when the agent does not meet their assumptions or does not solve their problem in the way they desire [6]. This may lead them to close the chat.

In short, research highlights that users' expectations shape the interaction experience with the chatbot, influencing their overall satisfaction. Expectations may revolve around the chatbot's capabilities of correctly interpreting the user's intent, providing timely information, and giving appropriate and relevant responses. In particular, expectations of receiving explanations about issues that are relevant to the user (e.g., unusual situations related to subscribed services) are certainly fundamental for customer care.

Some approaches related to users' expectations focus on the development of systems that generate either more accurate chatbot's clarifying questions (similarly to conversational search systems, e.g., [9]), or tailored responses aimed at anticipating users' information needs, or intentions. It is worth pointing out that the term "intention" here does not denote the user's "purchase intention" (see [10]), nor the user's "intent," a term that in task-oriented dialogue systems specifically indicates a user's goal expressed in an utterance.

A recent work considered the application of end-to-end task-completion neural DSs for the specific task of booking cinema tickets [11]. The authors also analyzed the impact of the errors of the natural language understanding module.

Di Lascio et al. [12] emphasized that the linguistic knowledge provided by users is often not sufficient to fulfill their expectations. We believe that in these cases the DS should use the knowledge on the domain context to produce a better interaction by predicting user intentions.

In this article, we propose a solution for situations in which users ask for an explanation of a certain unusual situation regarding the services that they have subscribed to, an issue that remained unexplored in previous research. We collected and analyzed a corpus of almost 3000 *customer conversations with a* DS using the log files of the DS of a telecommunication company (called COM-DS) and then selected the conversations where customers ask the DS for explanations. We, thus, found that about 5% of the total number of conversations from the corpus involves requests for explanations, and 50% of this specific type of conversations is about additional or unexpected charges on the customers' telephone accounts. This is a particularly dangerous situation for the company, as this kind of issue may result in customer churn.

Leveraging content selection mechanisms and Natural Language Generation (NLG) techniques, a new DS (called GEN-DS) was developed to adequately address vague or ungrammatical requests for explanations regarding unrecognized charges on the users' telephone accounts. Unlike other systems such as [11], GEN-DS does not rely on a grammatical and comprehensible linguistic input.

GEN-DS considers the history of the transactions on the user phone balance and discriminates between transactions that are typical and transactions that are uncommon. Furthermore, the developed model can distinguish between transactions that have a relevant impact on users' phone account and those that are negligible from the economic point of view. An important feature in the design and construction of GEN-DS is that, when asked about an unrecognized charge on the phone account, it produces a synthetic and useful answer for the user, as opposed to a more straightforward, but long-to-read and less immediate, response listing all the recent transactions on the user's account.

Our belief is that the developed techniques can be useful in other contexts, whenever a user observes some unusual state of a product, service, or system and asks a DS for clarifications about it. For example, the same kind of solution can be employed for managing the point account of frequent flyers of an airline, when users observe some anomaly in their point balance and ask for an explanation. The same goes for accumulated discounts in stores, or whenever there is a need to manage a balance of points, money, or whatever is affected by customer-company transactions.

The rest of this article is organized as follows. In Section II, we describe the corpus development, the design and implementation of GEN-DS, the design and execution of an experimentation with humans to evaluate GEN-DS. In Section III, we discuss the results of the experimentation. Finally, Section IV concludes this article.

II. Materials and Methods

This section provides the main contributions of this article. In Section II-A, we describe the construction of a dialogue corpus concerning explanation requests in the customer-care domain. In Section II-B, we provide formalization of the *evidence* notion to the customer transaction domain. In Section II-C, we describe GEN-DS, a DS specifically designed for managing the request emerged from corpus analysis and that produces answers based on evidence. In Section II-D, we provide a methodology for building experimental scenarios starting from real dialogues. Using these scenarios, in Section II-E, we describe an experimentation with users designed to evaluate the quality impact of GEN-DS for the case of explanation requests in the customer care domain. Finally, in Section II-F, we provide the results of the experimentations.

A. Building a Corpus of Explanation Requests in Customer-Care Dialogues Domain

For the purposes of this article, we first collected a sample corpus of conversations in Italian to find an empirical basis of our working hypothesis [13]. The supplementary material contains the corpus annotation. Due to the corporate privacy policy, the actual content was made available to the authors for the sole purpose of this research and cannot be publicly released.

The primary goal of the corpus analysis was to identify the main characteristics of these interactions, in terms of basic features—such as average number of turns per conversation and average turn length per user/agent—and to verify whether recurring linguistic behaviors could be found in customers' requests for explanations.

As mentioned in Section I, the dataset consists of real dialogues between customers and the DS of an Italian telecommunication company (COM-DS). This was created by selecting from a sample held over 24 hours a reduced subset that included requests for explanations from customers, thus using a simple string-matching method that extracted all conversations where the strings *perché* ("why", along with its orthographical variations, e.g., *perchè* or *xkè*) and *come mai* ("how come") occurred in the users' messages. The resulting corpus consists of 142 dialogues, for a total amount of 1540 turns, where each turn consists of one or more messages from one party. The collection features an average of about 11 turns per dialogue, and an average length of 9 tokens in customer turns and 38 tokens in the agent turns. The average dialogue length reported for this corpus is not in line with past literature that showed how task-oriented dialogues are typically shorter [27]. This could be explained by the fact

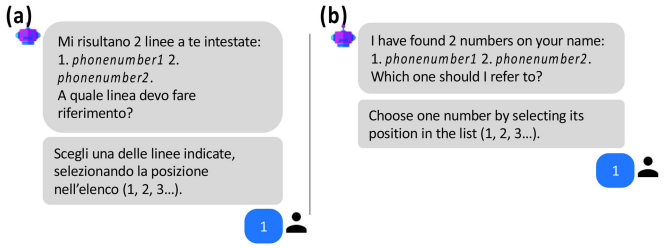


Fig. 1. Example of DS-customer interaction, with longer DS turns as opposed to more concise responses by the customer. On the left, the original dialogue excerpt in Italian, on the right its English translation.

that, especially in the case of the DS, a single turn most often consists of more than one message. Another striking difference is the average length of customers' messages compared to the ones from the chatbot. This difference is also due to the way the agent responses are currently structured; as a matter of fact, they usually include detailed information (for example, on invoice items or available options); conversely, customers' messages are generally more concise. In some cases, the latter are basic yes/no answers, or digits (1, 2, …) corresponding to the options provided by the agent in its previous message (see the example in Fig. 1). These divergences often lead to loops in communication in which both the user and the chatbot find themselves repeating the same requests or statements several times, within the same conversation, or providing irrelevant information that do not contribute to achieving the goal of the conversation, which is to provide the user with a clear and exhaustive explanation.

The architecture of many commercial DSs relies on the assumption that some relevant information is provided by the user utterance [5]. However, the exploratory analysis of this corpus proved that this assumption is sometimes false or only partially true. In fact, linguistic input in users' messages can be vague (see Example 1 below—we provide a rough English translation trying to replicate the vagueness and ungrammatical nature of the Italian original utterances), sometimes ungrammatical or not easy to follow (Examples 2 and 3), or too long and confusing (Example 4). In all such cases, the dialogue manager might need to ask for additional clarifications or to access some contextual information to compensate the lack of linguistic information.

1) *Perché mi sono stati scalati dei soldi.*
   (Why has some money been deducted (from my account)).
2) *Salve. Vorrei sapere perché ho pagato 0,50 cent. Per sms se li ho gratis E i 2,00 euro in piu per che cosa sono Grazie.*
   (Hello, I would like to know why I paid 0.50 cents. For texts if I have them for free And what about the additional 2.00 euros for what are they for Thanks.)
3) *Bg come mai mi è addebitato altri euro ho qualche cosa attivato a pagamento.*
   (GM how come I was charged other euros I have something activated for a fee.)
4) *Come mai mi vengono addebitati costi di <serviceName> quando non è stato mai richiesto da me E come mai la bolletta è passata da 36 a 57 euro Ho già disdetto <serviceName> dai cellulari, mi sa che devo dare disdetta anche dal fisso poichè mi sento costantemente vessato e truffato dalla vostra compagnia. Inutile dire che è praticamente impossibile parlare con un operatore al telefono. Vergogna.*
   (Why am I being charged for <serviceName> when it has never been requested for And why the bill has gone from 36 to 57 euros I have already canceled <serviceName> from mobile phones, I guess, I will have to cancel it from the landline as well because I constantly feel harassed and cheated by your company. Needless to say that it is impossible to speak to an operator on the phone. Shame on you.)
5) *Scusami ma vorrei sapere come mai mi vengono fatti certi addebiti.*
   (Sorry but I would like to know why there are some charges.)
6) *Salve vorrei sapere perchè mi sono stati presi 12€ invece che dieci dall'ultima ricarica.*
   (Hi, I would like to know why you charged 12€ instead of ten since last top-up.)
7) *Buongiorno, vorrei sapere perché ho il credito in negativo, nonostante abbia fatto una ricarica da 15€ proprio stamattina.*
   (Good morning, I would like to know why I have a negative balance, despite I made a 15€ top-up just this morning)

Prior to the GEN-DS design and development, some annotation experiments were carried out on the corpus, with the aim to explore possible recurring patterns underlying the user-chatbot interactions (see [13] for a more detailed description of the annotation scheme). In this process, we also observed that users' explanation requests typically fall under three main categories of request, that we briefly define as follows.

Category I: (58% of the occurrences in the corpus) a charge in the account is claimed, but no further information is provided (see Examples 1, 3, and 5).

Category II: (31% of the occurrences) the customer asks for an explanation about a charge providing vague information (Examples 2, 4, and 6).

Category III: (11% of the occurrences) the customer asks for an explanation about a negative balance (Example 7).

The corpus analysis, thus, provided evidence of the fact that in customer care interactions user requests can be vague and not informative enough, and they can be identified with (at least) one of the major categories we described previously. Considering this, we designed a new DS based on standard symbolic NLG techniques exploiting domain knowledge (see Section II-B) which can produce a response to the request (see Section II-C). We evaluated this DS based on the three categories identified in the corpus (see Section II-D).

B. Importance, Effect, and Evidence in Relational Domain-Context Knowledge

The need to connect domain-specific data to factual linguistic explanations has drawn much attention in the recent past [15]. A key role in this task is played by *content selection*, which determines what kind of information should be communicated

TABLE I
DC-KNOWLEDGE FOR EXAMPLES DC-K-1, DC-K-2, AND DC-K-3. EACH ROW INDICATES THE TRANSACTIONS OF A SPECIFIC CATEGORY. WE ASSUME THAT ALL THE TRANSACTIONS ON THE USER'S ACCOUNT ARE KNOWN

DC-K-1	M_1	M_2	M_3	M_4	M_5	M_6	M_7
S_1	9.99	9.99	9.99	9.99	9.99	9.99	9.99
S_2	0	0	0	0	2	2	2, 2
S_3	0	0	0	0	0	0	1.59

DC-K-2	M_1	M_2	M_3	M_4	M_5	M_6	M_7
S_1	10	10	10	10	10	10	10
S_2	0	0	0	0	0	2	2

DC-K-3	M_1	M_2	M_3	M_4	M_5	M_6	M_7
S_1	13	13	13	15	15	15	15
S_2	0	0	0	0	0	0	0.9,0.9, 0.9,0.9
S_3	0	0	0	0	0	0	1.99

to the user. Symbolic, statistical, and neural approaches have been proposed for this task (see [16] for a recent neural approach reporting a detailed survey on the state of the art).

In this article, we adapt the approach to content selection proposed by Biran and McKeown [17], by formalizing the notions of *effect*, *importance* and *evidence* to the specific context of our study, i.e., the customers' transactions stored in a relational database. We define the latter as *domain context knowledge* (DC-knowledge henceforth, cf., Table I). The original proposal in Biran and McKeown's work considers statistical classifiers based on linear discriminant functions, as linear SVMs. The notion of *effect* is anchored to the weight of a feature in the classification of a single data instance into a class y. In contrast, the notion of *importance* is anchored to the weight of a feature in the classification of all instances of the training set into a class y. The authors propose to combine these two notions in one single notion called *evidence.* They show that evidence can be used with the aim to select and order the features that should be communicated to the users, in an NLG system, for describing the trends of financial stock prices. In particular, importance values are narrative roles that "*… represent semantically clear concepts that non-experts readily understand and are rooted in the true details of the prediction*" [17, p. 1493]. Two roles played by a feature correspond to normal and exceptional evidence. Normal evidence is the case of a feature that is relevant both in the training set classifications (high importance) and in the current classification (high effect). In contrast, exceptional evidence is the case of a feature that is not relevant in the training set classifications (low importance) but is relevant in the current classification (high effect). The evidence model is defined only for statistical classifiers, and it is based on the existence of a training phase for defining importance and a classification phase for defining effect. Therefore, an application of this model to other settings needs a new definition of these two notions.

In GEN-DS we reformulate these notions for relational knowledge, that is typical of several applicative domains. Our original contribution is to use the evidence for giving priority to a specific transaction. The *importance* reflects the past relevance of a transaction, while the *effect* evaluates the current relevance of a transaction. The combination of these two notions determines the narrative role of a transaction, i.e., the transaction *evidence*. The importance sets out a sort of "expectation" for a transaction in contrast to the effect, which, if it does not match the importance, results in a "surprise" that is worth mentioning. Thus, a transaction has the narrative role of exceptional evidence in the case of low importance and high effect. A key idea in GEN-DS is that normal evidence is not surprising and should not be mentioned, at least primarily, in the generated message. In contrast, the cases of exceptional evidence are surprising and should be mentioned prominently.

It is worth pointing out that the two most important elements in this specific context are money and time. Therefore, we formalize our intuitions that a) the importance of customer-care service of a telecommunication company can be associated with the amount of money that the user usually spends for such service, and b) that its effect can be associated with the amount of money that the user spent for the service in the last month [18]. Formally, a transaction is a money transfer operation between a customer and the company (i.e., an amount paid for a certain service). As a result, each transaction sequence represents the different amounts paid along a time period for a specific service (transaction type). We, thus, define the *importance of a transaction sequence* as the mean of the normalized values of the transactions in the past K months. In the following examples, we consider the previous six months ($K = 6$). This value has been decided based on two considerations: a history going too deep in the past would generate messages that would be too verbose for the users and, moreover, from corpus analysis, it emerges that most of user requests do not concern very old transactions.

We define the *effect of a transaction sequence* as the normalized value of the transactions in the current month ($(K+1)^{th}$ month). Normalization is carried out by dividing the amount of the transactions by the maximum amount that the user has paid for that transaction. More formally, if S_i denotes the transaction sequence, M_j are the months, and T_{ij} are the transactions in the transaction sequence S_i occurred in month M_j, we can write Importance (1) and Effect (2) as

$$\text{Importance}(S_i) = \frac{1}{K} \frac{\sum_{j=1,\ldots,K} T_{ij}}{\max_{j=1,\ldots,K+1} T_{ij}} \tag{1}$$

$$\text{Effect}(S_i) = \frac{\sum_{s \in T_{iK+1}} s}{\max_{j=1,\ldots,K+1} T_{ij}}. \tag{2}$$

These numeric real values need to be discretized to classify importance and effect as *high* or *low*. In accordance with the original model in [17], we determine the smallest subset H of transaction sequences such that the sum of their importance/effect values is at least a fraction t of the total importance/effect. When such a subset is not unique, we consider the union of all the smallest subsets. Note that the value of t is a tunable value that should be empirically validated on the specific domain. In the following, as in [17], we use $t = 75\%$. We now describe three examples of DC-knowledge to illustrate these notions.

Example DC-K-1: The first DC-knowledge in Table I has three transaction sequences: S_1, with an amount of 9.99 euros (M_1–M_7), S_2 with an amount of 2 euros (M_5–M_7, appearing twice in M_7), and S_3 with an amount of 1.59 euros (M_7). From this data, we calculate importance and effect for S_1, S_2, and S_3, and their narrative roles. The importance of S_1 is $\text{Importance}(S_1) = \frac{1}{6} \frac{9.99+9.99+9.99+9.99+9.99+9.99}{9.99} = 1$. The importance of S_2 is $\text{Importance}(S_2) = \frac{1}{6} \frac{2+2}{2} = 0.33$. The importance of S_3 is $\text{Importance}(S_3) = \frac{1}{6} \frac{0}{1.59} = 0$. So, the sum of the importance values is 1.33 and its 75% is 1. The smallest subset H_I such that the sum of the importance values is at least 1 is $H_I = \{S_1\}$, so S_1 has high importance, while S_2 and S_3 have low importance.

The effect of a transaction sequence is given by the values in the current month, so the effect of S_1 is $\text{Effect}(S_1) = \frac{9.99}{9.99} = 1$, the effect of S_2 is $\text{Effect}(S_2) = \frac{2+2}{2} = 2$, and the effect of S_3 is $\text{Effect}(S_3) = \frac{1.59}{1.59} = 1$. The sum of the effect values is 4 and its 75% is 3. The smallest subset H_E such that the sum of the effect is at least 3 is $H_E = \{S_1, S_2, S_3\}$, hence, S_1, S_2, and S_3 all have high effect. As a result, combining the discrete values of importance and effect, S_1 is normal evidence, and S_2 and S_3 are both exceptional evidence.

Example DC-K-2: This example of DC-knowledge (the second example in Table I) has two transaction sequences: S_1, with an amount of 10 euros (M_1–M_7), and S_2 with an amount of 2 euros (M_6–M_7). Also from this data, we calculate importance and effect for S_1 and S_2. The importance of S_1 is $\text{Importance}(S_1) = \frac{1}{6} \frac{10+10+10+10+10+10}{10} = 1$. The importance of S_2 is $\text{Importance}(S_2) = \frac{1}{6} \frac{2}{2} = 0.17$. So, the sum of the importance values is 1.17 and its 75% is 0.88. The smallest subset H_I such that the sum of the importance values is at least 0.88 is $H_I = \{S_1\}$, so S_1 has high importance, while S_2 has low importance. The effect S_1 is $\text{Effect}(S_1) = \frac{10}{10} = 1$, and the effect of S_2 is $\text{Effect}(S_2) = \frac{2}{2} = 1$. The sum of the effect values is 2 and its 75% is 1.5. The smallest subset H_E such that the sum of the effect is at least 1.5 is $H_E = \{S_1, S_2\}$, hence, S_1 and S_2 have high effects. As a result, combining the discrete values of importance and effect, S_1 is normal evidence, and S_2 is exceptional evidence.

Example DC-K-3: This example of DC-knowledge (see the third example in Table I) has three transaction sequences: S_1, with amounts of 13 euros (M_1–M_3) and 15 euros (M_4–M_7), S_2 with an amount of 0.9 euros (four times in M_7), and S_3 with an amount of 1.99 euros (in M_7). The importance of S_1 is $\text{Importance}(S_1) = \frac{1}{6} \frac{13+13+13+15+15+15}{15} = 0.94$, while $\text{Importance}(S_2) = \text{Importance}(S_3) = 0$. The sum of the importance values is 0.94 and its 75% is 0.71. The smallest subset H_I such that the sum of the importance values is at least 0.71 is $H_I = \{S_1\}$, so S_1 has high importance, while S_2 and S_3 have low importance.

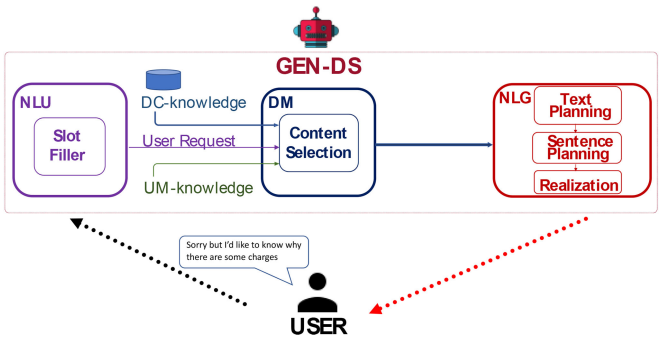


Fig. 2. Architecture of GEN-DS.

The effect of S_1 and S_3 is 1, while the effect of S_2 is $\text{Effect}(S_2) = \frac{0.9+0.9+0.9+0.9}{0.9} = 4$. The sum of the effect values is 6 and its 75% is 4.5. The smallest subset H_E such that the sum of the effect is at least 4.5 can be $H_E = \{S_1, S_2\}$ or $H_E = \{S_2, S_3\}$. The subset H_E is the union of the two cases, i.e., $H_E = \{S_1, S_2, S_3\}$, hence, S_1, S_2, and S_3 have high effect. Thus, S_1 is normal evidence, and S_2 and S_3 are exceptional evidence.

In the following section, we describe how our reformulation of evidence can be used to guide the content selection process in GEN-DS.

C. Designing and Implementing GEN-DS

GEN-DS follows the classical cascade architecture depicted in Fig. 2 [5]. GEN-DS is composed of three modules, which are *natural language understanding* (NLU), *natural language generation* (NLG), and *dialogue manager* (DM). The NLU module is devoted to the interpretation of the user's utterances. The NLG module is devoted to the final generation of the DS answer. The DM, that takes the input from the NLU module and produces the output for the NLG module, is devoted to managing all semantic and pragmatic elements that influence the future development of the dialogue. For instance, the DM can decide to answer with a question to a question. This classical architecture has a long history, but several advancements have recently been adopted. For instance, most modern DSs use NLU techniques based on machine learning to fill the important conceptual slots (e.g., *intents* and *entities*, [11], [19]) of the domain. Moreover, recent developments of neural NLG can be adopted also in some specific cases of generation in dialogues [20]. However, apart from systems devoted to *chit-chat*, also in modern task oriented DSs all the information must be coordinated by the DM to update the internal state of the DS and to produce the next dialogue act [5].

Many DSs assume that a relevant part of the necessary information is provided by the user's utterance, analyzed by the NLU module, and passed to the DM as *User Request* in Fig. 2 [5]. However, as outlined in Sections I and II-A, this assumption is only partially true in customer-care domain. Even a very advanced NLU module cannot make a detailed analysis in the case

of a vague request as the one in Example 5 (see Section II-A). Indeed, in this case very often commercial DSs apologize and ask users to repeat their request with more details [9]. Moreover, some user utterances are ungrammatical, as Example 3, and cannot be analyzed at all.

To provide better responses, in the case of vague or ungrammatical user requests, GEN-DS can resort to two other sources of information: the domain context knowledge (*DC-knowledge*) and the user model knowledge (*UM-knowledge*). In particular, the GEN-DS system depicted in Fig. 2 has been designed for overcoming the limitations of the *apologize-and-ask-to-repeat* strategy by using an NLG approach that exploits the DC-knowledge. In accordance with other systems developed for other domains [14], in GEN-DS the DC-knowledge plays a central role to produce the content of the answer: the basic idea is to produce responses that have exceptional evidence with respect to the specific DC-knowledge. In this way, we can anticipate the user intention by building interesting and concise answers to vague or ungrammatical requests. So, in GEN-DS, we specifically designed the NLU, the DM and the NLG modules for managing the explanation requests found in the corpus, but we believe that GEN-DS could be easily integrated in more general DSs designed for generic interactions. Note that the other source of information, that is the UM-knowledge, is not used for content selection but it plays a role in the realization submodule deciding the linguistic details of the generated sentence. In the remaining part of this section, we describe the main features of the NLU, DM, and NLG modules of GEN-DS.

NLU Module: It is based on regular expressions and is inspired by the NLU features of the COM-DS. In particular, the NLU distinguishes between the cases of generic or charge-related explanation requests. In some sense, the NLU must fill a single semantic slot related to the type of the request. For instance, in the case of ungrammatical utterances such as the one in Example 3, NLU returns a generic user request, while in the utterance from Example 2, the NLU returns a specific user request.

DM Module: It deals with the content selection task, that implements the notion of evidence formalized in the customer-care domain of a telecommunication company (see Section II-B), by providing the NLG module with all the transactions with their values of evidence. Moreover, in some specific cases, even the value of the user balance and the value of the total charges are provided to the NLG module. For instance, in the case of Example DC-K-1 in Table I, the content selection will pass the total amount of the charges, the transactions S_1, S_2, S_3 together with the information that S_2 and S_3 are both cases of exceptional evidence, and S_1 is normal evidence.

NLG Module: It includes, in turn, the submodules in charge of the three typical steps that characterize symbolic approaches to NLG, i.e., text planning, sentence planning, and realization [21].

In general, text planning for NLG concerns both the selection of the salient information and its organization in a causal and temporal structure [21]. Indeed, since content selection is managed by the DM, the role of text planning in GEN-DS is to order the information provided the DM. We designed a very simple text planning schema to sort the content in a specific ordered list,

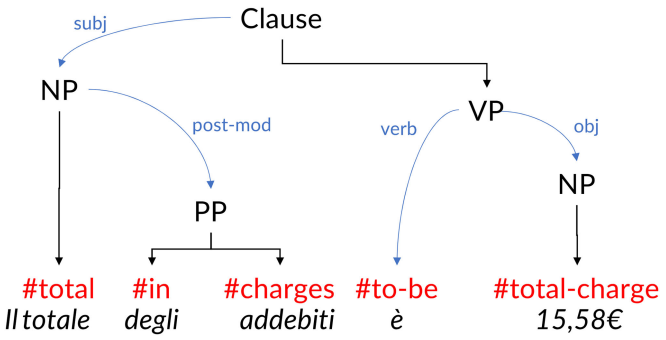


Fig. 3. Syntactic template for a declarative sentence. The leaves of the tree (in red, starting with #) contain lexical items that will be instantiated by the realizer.

that will be used in the realization for ordering the sentences: 1) information on user balance, 2) information on total charges, 3) information on the transactions with exceptional evidence, and 4) information on the transactions with normal evidence.

The sentence planner of GEN-DS is a rule-based module that defines the number and the types (e.g., passive, declarative) of the sentences in the final message, defines which sentences need to be merged for fluency, and defines which lexical elements to use for each sentence. GEN-DS uses a sentence planner previously adopted in several applicative projects of data-to-text generation [22]. The syntactic information on the sentences is encoded in a few predefined syntactic templates: Fig. 3 shows the syntactic template used to generate the first sentence in Example 8 (see below). The syntactic template is an unordered tree encoding notions from both constituency and dependency theories of syntax: it adopts both *phrases* from constituency theory (e.g., Noun Phrase, NP, Verbal Phrase, VP) and *relations* from dependency theory (such as subject and object, abbreviated to *subj* and *obj* in Fig. 3). Note that the trees do not fully specify the word inflection and the word order. The leaves in Fig. 3 (starting with #) indicate lexical items that will be specified in the realization by using the corresponding numeric values and the realizer domain dictionary.

The sentence planner decides which templates to use following two principles. The first principle regards visual readability: the sentences will be read in a textual chat; thus, shorter sentences are preferable. The second principle regards linguistic fluency: it prescribes to aggregate information on transactions, which have the same value of evidence. For instance, in the case of Example DC-K-1 in Table I, S_1 will be communicated in one single sentence, and S_2 together with S_3 in another single sentence.

Finally, the realization process is implemented by using the SimpleNLG-IT library [23], which completes the syntactic templates with the necessary morpho-syntactic and orthographic information of the Italian language and the correct numeric values. SimpleNLG-IT is a rule-based realizer that formalizes the Italian grammar by producing morphologically correct word inflections and word orders. In the current implementation of GEN-DS, SimpleNLG-IT is the only module that uses information provided by the user model: for young users (less than

twenty years), the pronoun *tu* is used (a colloquial second person pronoun), in contrast to the pronoun *lei* used for older people (a more formal second person pronoun).

Given Example DC-K-1 reported in Table I, the final output produced by SimpleNLG-IT is the one shown in Example 8.

8) *Il totale degli addebiti è 15,58€. Recentemente hai pagato 4,00€ (2x€2,00) per l'Offerta Base Mobile e 1,59 € per l'Opzione ChiChiama e Richiama. Infine, come al solito, hai pagato il rinnovo dell'Offerta 20 GB Mobile (€9,99).* (The total charge is €15.58. Recently, you paid €4.00 (2x€2.00) for the Basic Mobile Plan and €1.59 for the WhoCalled and CallMeBack Options. In addition, as usual, you paid for the renewal of the 20 GB Mobile Plan (€9.99).)

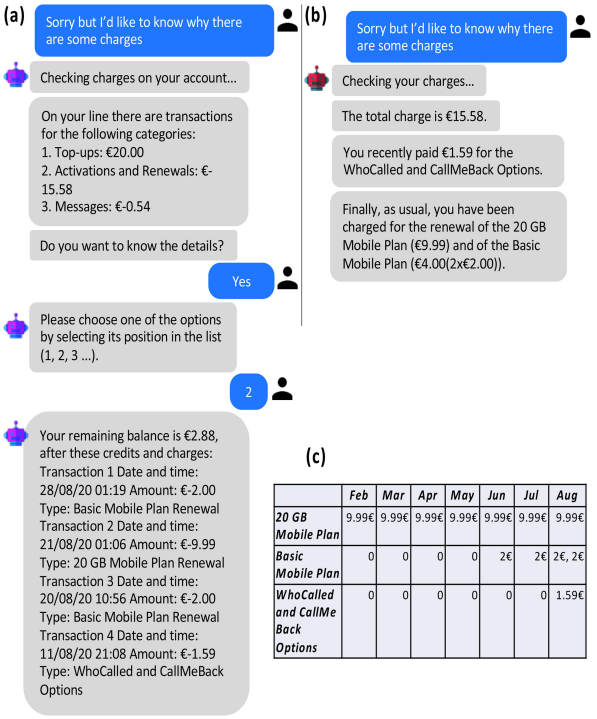
D. Building Experimental Scenarios

In this section and in the next, we present the first experimental human-based evaluation of GEN-DS.

Our main goal was to design a realistic experimentation for GEN-DS. Different methodologies and frameworks have been proposed over the years to evaluate DSs [24]–[26]. However, we observed in our sample corpus that, among the possible situations where a customer may ask for assistance, just a small percentage (5%, see Section II-A) include explanation requests, which are the focus of our research. Therefore, it would not be trivial to collect enough samples in an interactive unconstrained experimentation conducted with live users. Thus, we decided to follow one of the experimental protocols defined in [27], where six prototypical dialogues and contexts, called *scenarios*, were generated offline and evaluated by users along several properties. Notice also that the corpus analysis (see Section II-A) showed that the whole range of possible customers' requests falls into three main categories. Using these three categories, we designed four distinct scenarios (reported in the supplementary material). In our study, a scenario is a prototypical situation consisting of a request for explanation, contained in the user utterance, and a specific DC-knowledge. It is worth pointing out that the four scenarios we devised ensure the coverage of all the categories extracted from the corpus and provide a good statistical power at the same time (see Section II-F). Moreover, the number of scenarios is comparable to the number of scenarios used in [27].

We built two scenarios (Scenario 1 and Scenario 3) using a linguistic input from Category 1 (the largest category), one scenario (Scenario 2) using a linguistic input from Category 2, and finally one scenario (Scenario 4) using a linguistic input from Category 3. For each scenario, we randomly extracted from the corpus one dialogue of the corresponding category. We then have used the user explanation request that opens the dialogue and the DC-knowledge of that dialogue to produce an answer with the GEN-DS system (see Fig. 4, with translated dialogues).

We recovered the DC-knowledge for the specific dialogue from the commercial system database: it consists of the user transactions of the last two months. As explained in Section II-B, we formalized the evidence-effect-importance over a period of seven months. We, thus, augmented the actual DC-knowledge



	Feb	Mar	Apr	May	Jun	Jul	Aug
20 GB Mobile Plan	9.99€	9.99€	9.99€	9.99€	9.99€	9.99€	9.99€
Basic Mobile Plan	0	0	0	0	2€	2€	2€, 2€
WhoCalled and CallMe Back Options	0	0	0	0	0	0	1.59€

Fig. 4. Scenario 1 used in the experimentation translated in English. (a) Original dialogue selected from the corpus between a user and COM-DS. (b) Dialogue generated by GEN-DS. (c) Common DC-knowledge. Due to space constraints, the original dialogue excerpt in Italian is provided in the supplementary material.

obtained by the commercial system with ad-hoc realistic additional knowledge on the previous five months. With the aim to evaluate the utility of our formalization of DC-knowledge in realistic but similar situations, we designed Scenarios 1 and 3 keeping the same user's request though varying the transactions in the augmented knowledge.

The scenarios generated using COM-DS have an average of about six turns per dialogue (with an average length of six tokens in customer turns and 46 tokens in the agent turns). In contrast, the responses generated by GEN-DS, with two turns per dialogue (but an average length of 55 tokens in customer turns and 15 tokens in the agent turns) help reduce the number of turns overall, as the system provides the users with the required information right after the explanation request.

E. Participants and Experimental Procedure

To validate the experimental hypothesis that users prefer dialogue systems where the answers are generated (selected and/or ordered) based on the evidence of the transactions, we prepared an online questionnaire. In line with previous work on similar tasks (such as [11], [20], [24], [27]), a pairwise comparison was carried out, in that users were asked to evaluate two different DSs: the original commercial system COM-DS, used to build the dialogue corpus (see Section II-A), and an implementation of GEN-DS (see Section II-C). We invited several colleagues, students, and acquaintances by email, asking for friendly participation without rewards. Around one hundred people have been invited and 54 users participated in our experiment. In

total, 30 users (55.6%) were students, 23 users (42.6%) were employees, and one user (1.8%) was a teacher. Most of the users (29 users, 53.7%) were 18–30 years old, 11 users (20.4%) were 31–45 years old, 13 users (24.1%) were 46–60 years old, and only one user (1.8%) was less than 18 years old. Finally, all the participants were Italian native speakers, and ten of them (18.5%) had no experience with DSs before this experimentation. Prior to participation we informed users that the survey concerned DSs, that no sensible data would be collected and that we ensured anonymity. As an introduction to the questionnaire, we informed users that they would be presented with different pairs of dialogues produced by different customer care DSs and that they would be asked to evaluate them over four different scenarios. In each pair of dialogues (see Fig. 4), the systems were simply tagged as System A and System B and arranged as Latin squares. We released the questionnaire as an online form, built with Google Form, composed of 43 questions, where 36 questions concerned dialogue scenarios, six questions concerned user profile information (age, educational qualification, occupation, technological skill, and their previous experience with chatbots), and one question was an open question for free comments.

We followed the experimental protocol defined by Demberg et al. [27]: in the questionnaire, for each scenario, we presented both the original dialogue extracted from the corpus and a dialogue generated by GEN-DS. Both dialogues have the same user explanation request and the same DC-knowledge. For each dialogue the users were asked to rate four specific properties of the DSs, that are *usefulness, necessity, understandability,* and *quickness*. Users were presented with a statement regarding such qualities and were asked to specify their agreement with a seven-point Likert scale where 1 corresponds to "I completely disagree" and 7 to "I completely agree" (the supplementary material contains the complete questionnaire). The statements are, respectively, as follows.

Usefulness: "*All the information provided by the system is USEFUL to respond to your request*" (Italian: "Le informazioni fornite dal sistema sono tutte UTILI per rispondere alla tua richiesta"). The rationale of this statement is to check if all the information provided by the DS concerns the fulfilment of the explanation request, that is there is no useless information provided in relation to the explanation request. To rate this statement, the user needs to consider the DS sentences in the dialogue together with the transaction table associated with the scenario, that is the specific DC-knowledge. In a sense, usefulness pertains to the notion of *precision* used in information extraction.

Necessity: "All the information NECESSARY to answer your request has been presented by the system" (Italian: "Tutte le informazioni NECESSARIE per rispondere alla tua richiesta sono state presentate dal sistema"). The rationale of this statement is to check whether the information contained in the specific scenario DC-knowledge provided by the DS concerns the fulfilment of the explanation request, that is whether there is no unnecessary information in relation to the explanation request. Also in this case the user needs to consider the DS sentences in the dialogue together with the specific scenario DC-knowledge. In a sense, necessity pertains to the notion of *recall* used in information extraction.

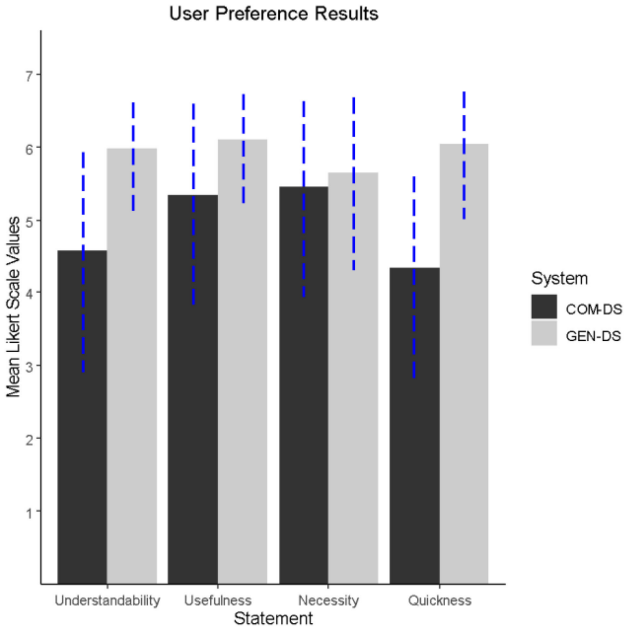


Fig. 5. Mean values and standard deviations for usefulness, necessity, understandability, and quickness for the two systems.

Understandability: "*The system provided the information in a way that it is easy to understand*" (Italian: "Il sistema ha fornito le informazioni in un modo facile da comprendere"). The rationale of this statement is to evaluate the comprehensibility of the language used by the DS in the conversation.

Quickness: "The system was quick in allowing you to find the salient information" (Italian: "Il sistema è stato rapido nel permetterti di trovare le informazioni salienti"). The rationale of this statement is to ask users to evaluate the "efficiency" of the text generated by the DS concerning the requested explanation in quickly obtaining the desired information. Notice that this property does not concern computational performance (e.g., response time) of the system, but just "textual" features such as conciseness. We provide this statement since this notion seems to be particularly important for the user satisfaction in DS interactions [27].

Satisfaction: "*Which system would you recommend to a friend?*" (Italian: "Quale dei due sistemi consiglieresti ad un amico?"). Following the evaluation schema proposed by Demberg et al. [27], to assess user satisfaction, in the questionnaire we include also a binary question asking whether the user prefers one system or the other.

F. Results

For each property presented previously, we discuss the results (see Fig. 5–the supplementary material contains the answers to the questionnaire given by the 54 users).

Usefulness: This statement assessed the user's confidence that all the information mentioned by the systems is relevant. GEN-DS ($M = 6.10$, $SD = 0.95$) had a higher mean compared to COM-DS ($M = 5.35$, $SD = 1.58$) and the difference was

TABLE II
CORRELATION BETWEEN SYSTEM PROPERTIES AND USER AGE

Property	**Understandability**		**Usefulness**		**Necessity**		**Quickness**	
System	COM-DS	GEN-DS	COM-DS	GEN-DS	COM-DS	GEN-DS	COM-DS	GEN-DS
Age	-.441**		-.305**		-.436**		-.371**	
	-.406**	-.193	-.295*	-.224	-.327*	-.344*	-.228	-.426**

*Indicates a significant correlation at the 0.05 level (2-tailed).
**Indicates a significant correlation at the 0.01 level (2-tailed).

significant ($t(215) = 6.16$, $p<0.001$) according to a two-tailed paired t-test.

Necessity: This statement assessed the user's confidence that all the relevant information in the DC-knowledge has been mentioned by the system in the dialogue. The evaluation seems to show a slight preference for the GEN-DS system ($M = 5.65$, $SD = 1.35$) with respect to COM-DS ($M = 5.45$, $SD = 1.51$). However, this preference is not statistically significant ($t(215) = 1.52$, $p = 0.07$).

Understandability: This statement assessed the user's confidence that all the information mentioned by the systems is comprehensible. The mean of the GEN-DS system was rated significantly higher ($M = 5.98$, $SD = 0.96$) on this statement in comparison to COM-DS ($M = 4.58$, $SD = 1.68$, $t(215) = 11.02$, $p<0.001$ according to a two-tailed paired t-test).

Quickness: This statement assessed the user's confidence that the system presents information quickly. The mean of the GEN-DS system was rated significantly higher ($M = 6.04$, $SD = 1.04$) on this statement in comparison to COM-DS ($M = 4.35$, $SD = 1.56$, $t(215) = 13.04$, $p<0.001$ according to a two-tailed paired t-test).

Satisfaction: A significant preference for the GEN-DS system was observed. From a total of 216 choices in the experiment (54 participants x 4 dialogue pairs), GEN-DS was preferred 157 times (72.7%), whereas the dialogue from the corpus was preferred only 59 times (27.3%). This difference is significant according to a two-tailed binomial test ($p<0.001$). Thus, the null hypothesis that the corpus-based system is preferred at least as GEN-DS can be rejected with high confidence. We conducted a post hoc power analysis to assess whether we had sufficient subjects, and we obtained a good power value (power = 1) for usefulness, understandability, quickness, and satisfaction.

G. Subgroup Analysis

As post hoc analysis, to search for relations between the features characterizing the users and the scores that the users gave to the system properties in the questionnaire, we computed correlations between subgroups of users and how these users rated system properties. In Table II, we report the Pearson correlation coefficients for numerical features and the Spearman's rank correlation coefficients for categorical features. The only feature that showed significant correlation with users' scores is the age of the users. We report the correlation between the users' ages and the scores assigned by the users to the four system properties. It is interesting to notice that all four properties show highly significant negative correlations with age when aggregating COM-DS and GEN-DS scores: in other words, this means that, as age grows, users tend to give lower scores regardless of the system. When the correlation is computed separately on COM-DS and GEN-DS, we have negative significant correlation values only for specific systems/properties. While the understandability of COM-DS decreases as age grows, this does not impact on GEN-DS. Indeed, we found that people in the range 18–30 gives 5.35/7 for understandability whilst people in the range 45–60 gives 3.98/7. Moreover, older users deem GEN-DS slower with respect to young users. Indeed, for the quickness property, users in the range 18–30 give 6.5/7 whilst those in the range 45–60 give 5.71/7. However, given the limited number of users in our experimentation, and the fact that more than half of the users are between 18 and 30 years old, these results should be replicated in a specifically designed experimentation.

III. DISCUSSION

In this section, we review the experiment results by considering the main research question underlying this article, that is whether we can improve dialogue systems in the case of low-quality linguistic input by using content selection techniques that predict users' intentions.

As observed in our corpus, users often ask for explanations without providing enough linguistic information. Most commercial DSs, as COM-DS, the DS considered in this paper, cannot properly manage these dialogues. We built the GEN-DS system to properly address this issue. In particular, by using the formalization of evidence, GEN-DS is able to provide a tentative answer to unclear questions. The experiment described in Section II-E asks users to compare, along several properties, the responses provided by COM-DS in real conversations with the responses generated by GEN-DS based on the same DC-knowledge.

The results reported in Fig. 5 show that users deem GEN-DS superior to COM-DS with respect to the properties of usefulness, understandability, and quickness. The users report that GEN-DS presents the same DC-knowledge in a way that is more useful, understandable, and quick with respect to COM-DS. All these three properties are related to the way in which the relevant information is organized in the dialogue. The higher values reported for usefulness and quickness confirm that GEN-DS provides more relevant information in a more concise way. As regards usefulness, such results can be attributed to the shorter dialogue length produced by the sentence planning module, which relies on two main principles (see Section II-C): one that promotes shorter sentences for the sake of readability, and one that aggregates the sentences based on the evidence model described in Section II-B, to improve linguistic fluency. The

higher results reported for understandability seem to confirm that combining these principles does not come at the expense of linguistic clarity, on the contrary, it enhances it. In contrast, GEN-DS and COM-DS are not deemed statistically different with respect to necessity, that is the property of a dialogue to present all the necessary contextual information. This was an expected result as COM-DS offers a detailed account of all the transactions of the last two months, and the information it provides is a superset of the information provided by GEN-DS. The overall preference of the users toward GEN-DS is confirmed by the satisfaction question, where we asked them to explicitly compare COM-DS and GEN-DS. In short, based on these findings, we can conclude that *we can improve DSs in the case of low-quality linguistic input by predicting the users' intentions.*

IV. Conclusion

In this paper, we proved that the quality of a DS in the domain of customer care can be improved by using the notion of evidence. Based on a preliminary corpus analysis, we observed that several explanation requests, that are often used by users to start their conversations with DSs, are vague or ungrammatical or, more in general, hardly understandable. In such cases, the dialogue manager does not have sufficient linguistic information to produce a meaningful answer. Most commercial DSs deal with this situation with a simple apologize-and-ask-to-repeat strategy. However, a possible handling strategy may consist in using the extra-linguistic knowledge related to the dialogue, such as the UM-knowledge (about the user) and the DC-knowledge (about the domain). In the specific context of customer care, the DC-knowledge consists of the commercial transactions between the user and the company. For many companies this kind of information is encoded in relational structures.

In this paper, we provided a formal definition of *evidence* for relational data that can be used to select content from the DC-knowledge, and we implemented this notion on a data-to-text generation system that we called GEN-DS.

The experimentation in Section II-E compared real dialogues taken from a corpus of human/COM-DS conversations for the customer care service of a telecommunication company with synthetic dialogues generated by the GEN-DS system. The questionnaire results showed preferences for GEN-DS dialogues by measuring different properties: usefulness, necessity, understandability, quickness, and general satisfaction. Thus, we showed that this formalization can improve the quality of the dialogues in the customer-care domain.

To test GEN-DS in the specific case of a conversation that starts with hardly understandable user sentences, we designed the experimentation as a simulated dialogue rather than a real one. We are aware of the limit of this kind of experiment, but the specificity of our research goal does not allow to design a natural interaction with users to judge the contribution of the notion of evidence for content selection.

We believe that the results of this article can be easily extended to other application domains. Indeed, the core idea of our research is the formalization of the notion of evidence for relational knowledge and its application to NLG in the customer care domain. This notion was originally defined by Biran and McKewon in the field of machine learning and, inspired by their formalization of evidence in terms of importance and effect, we proposed in this paper 1) to encode the *past* behavior of the system into *importance*, 2) to encode the *recent* behavior of the system into *effect*. So, to apply our approach of generation based on evidence in DC-knowledge, one needs to reformulate the notions of evidence, importance, and effect for the specific task related to the DS.

Finally, a central question that arises from our research concerns the possibility of mixing together the DC-knowledge, the UM-knowledge, and the user requests. As a future work, it could be interesting to investigate how the notion of evidence can be used also in case of *understandable* linguistic input and how the user model can contribute to this process.

References

[1] J. Grudin and R. Jacques, "Chatbots, humbots, and the quest for artificial general intelligence," in *Proc. Conf. Hum. Factors Comput. Syst.*, ACM, 2019, pp. 1–11.

[2] M. Chung, E. Ko, H. Joung, and S. J. Kim, "Chatbot e-service and customer satisfaction regarding luxury brands," *J. Bus. Res.*, vol. 117, pp. 587–595, 2018.

[3] E. Kimani, K. Rowan, D. McDuff, M. Czerwinski, and G. Mark, "A conversational agent in support of productivity and wellbeing at work," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interaction*, 2019, pp. 1–7.

[4] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots," *Int. J. Human-Comput. Stud.*, vol. 151, Jul. 2021, Art. no. 102630.

[5] M. McTear, Z. Callejas, and D. Griol. *The Conversational Interface: Talking to Smart Devices*. 1st ed. Berlin, Germany: Springer, 2016.

[6] M. Jain, P. Kumar, R. Kota, and S. N. Patel, "Evaluating and informing the design of chatbots," in *Proc. Designing Interactive Syst. Conf.*, 2018, pp. 895–906.

[7] A. Følstad and M. Skjuve, "Chatbots for customer service: User experience and motivation," in *Proc. 1st Int. Conf. Conversational User Interfaces*, 2019, pp. 1–9.

[8] A. Følstad, M. Skjuve, and P. B. Brandtzaeg, "Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design," in *Proc. INSCI*, 2018, pp. 145–156.

[9] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft, "Asking clarifying questions in open-domain information-seeking conversations," in *Proc. ACM SIGIR*, 2019, pp. 475–484.

[10] K. Fang, Q. Zhang, Z. Zhuang, and Z. Zhang, "Making recommendations better: The role of user online purchase intention identification," in *Proc. Int. Conf. Softw. Netw.*, 2016, pp. 1–4.

[11] X. Li, Y.-N. Chen, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 733–743.

[12] M. Di Lascio et al., "Natural language generation in dialogue systems for customer care," in *Proc. CLiC-it*, CEUR, 2020, pp. 1–6.

[13] M. Sanguinetti, A. Mazzei, V. Patti, M. Scalerandi, D. Mana, and R. Simeoni, "Annotating errors and emotions in human-chatbot interactions in Italian," in *Proc. 14th Linguistic Annotation Workshop*, ACL, 2020, pp. 1–12.

[14] G. Stoilos, S. Wartak, D. Juric, J. Moore, and M. Khodadadi, "An ontology-based interactive system for understanding user queries," in *Proc. Eur. Semantic Web Conf.*, 2019, pp. 330–345.

[15] E. Reiter, "Natural language generation challenges for explainable AI," in *Proc. NL4XAI*, 2019, pp. 3–7.

[16] R. Puduppully and M. Lapata, "Data-to-text generation with macro planning," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 510–527, 2021.

[17] O. Biran and K. McKeown, "Human-centric justification of machine learning predictions," in *Proc Int. Joint Conf. Artif. Intell.*, 2017, pp. 1461–1467.

[18] L. Anselma, M. Di Lascio, D. Mana, A. Mazzei, and M. Sanguinetti, "Content selection for explanation requests in customer-care domain," in *Proc. INLT4XAI*, 2020, pp. 5–10.

[19] A. M. Preininger et al., "Artificial intelligence-based conversational agent to support medication prescribing," *JAMIA Open*, vol. 3, no. 2, pp. 225–232, 2020.
[20] T. Zhao, A. Lu, K. Lee, and M. Eskenazi, "Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability," in *Proc. SIGDIAL*, 2017, pp. 27–36.
[21] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, no. 1, pp. 65–170, 2018.
[22] L. Anselma and A. Mazzei, "Building a persuasive virtual dietitian," *Inform.*, vol. 7, no. 3, 2020, pp. 1–26.
[23] A. Mazzei, C. Battaglino, and C. Bosco, "SimpleNLG-IT: Adapting SimpleNLG to Italian," in *Proc. Int. Natural Lang. Gener. Conf.*, 2016, pp. 184–192.
[24] M. Danieli and E. Gerbino, "Metrics for evaluating dialogue strategies in a spoken language system," in *Proc. AAAI Symp. Empirical Methods Discourse Interpretation Gener.*, 1995, pp. 34–39.
[25] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with PARADISE," *Natural Lang. Eng.*, vol. 6, no. 3/4, pp. 363–377, 2000.
[26] J. Deriu et al., "Survey on evaluation methods for dialogue systems," *Artif. Intell. Rev.*, vol. 54, pp. 755–810, 2021.
[27] V. Demberg, A. Winterboer, and J. D. Moore, "A strategy for information presentation in spoken dialog systems," *Comput. Linguistics*, vol. 37, no. 3, pp. 489–539, 2011.

Alessandro Mazzei received the Graduate degree in physics from the University Federico II of Naples, Naples, Italy, in 2000, and the Ph.D. degree in computer science from the University of Turin, Turin, Italy, in 2005.

He has been an Assistant Professor in Computer Science with the University of Turin since 2006. His main research interests include the areas of Parsing, Natural Language Generation, and Dialogue systems. He has authored more than 90 papers published in international journals and international refereed conferences.

Luca Anselma received the Ph.D. in computer science from the University of Turin, Turin, Italy, in 2006.

Since 2006, he has been an Assistant Professor in computer science with the University of Turin. He has authored more than 60 papers published in international journals, books, and international refereed conferences. His main research interests include the areas of temporal reasoning, temporal databases, and medical informatics.

Manuela Sanguinetti received the Ph.D. degree in computer science from the University of Turin, Turin, Italy, in 2016.

She is currently a Postdoctoral Researcher with the University of Cagliari. Her research interests include linguistic annotation, dependency syntax, text processing of user-generated content, and knowledge-based question answering.

Amon Rapp received the Ph.D. degree in sciences of language and communication from the University of Torino, Turin, Italy, in 2015.

He is an Assistant Professor with the Department of Computer Science, University of Torino. His research interests include self-tracking and wearable devices, behavior change technologies, intelligent agents, and video games.

Dario Mana has been with the Telecom Italia Mobile, Rome, Italy, since 2001, received the M.Sc. degree in software engineering from the Politecnico di Torino, Turin, Italy, in 2000. He was involved in many innovation and engineering projects regarding social networks, network protocols, interactive TV, semantic technologies, and distributed systems. His research interests include natural language processing and virtual assistants.

Md. Murad Hossain received the B.Sc. and M.Sc. degrees in statistics from the Jahangirnagar University, Savar, Bangladesh, in 2010 and 2011, respectively. He is currently working toward the Ph.D. degree in modeling and data science with the Department of Modeling and Data Science, University of Turin, Turin, Italy.

He is currently an Assistant Professor (on leave) with the Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh. His research interests include NLP, statistics, data science, machine learning, and public health.

Viviana Patti received the master's degree in philosophy from Università degli Studi di Torino, Turin, Italy, in 1996, and the Ph.D. degree in computer science from Università degli Studi di Torino, Turin, Italy, in 2002. He is currently an Associate Professor of computer science with the University of Turin, Turin, Italy, and part of the scientific board and executive committee of the Center for Logic, Language, and Cognition. Her recent research interests include the intersection of computational linguistics and affective computing, with a focus on sentiment analysis, irony detection and hate speech detection on social media, and a specific interest on NLP for social good.

Rossana Simeoni received the master's degree in computer science from Università degli Studi di Torino, Turin, Italy, in 1991.

Since 1993, she has been applying her competences in Telco R&D departments. She is a Innovation Project Manager with Telecom Italia Mobile, Rome, Italy, and Adjunct Professor with the University of Torino, Turin, Italy. Her research interests include interaction design, intelligent interactive systems, conversational agent, and natural language processing.

Lucia Longo received the master's degree in psychology from Università Cattolica del Sacro Cuore, Milano, in 2007.

She was a Psychoterapist in 2012. She has been UX expert at Telecom Italia Mobile, Rome, Italy, since 2014. Her research interests include user experience and ergonomic aspects in several fields, among which AR/VR and mixed reality, AI - conversational agent, and NLP, from a user centered design perspective.

Appendix C

Related code repository of thesis

C.1 Related code repository link

You can find all the codes at: GitHub - Reasoning-NLG-Unito/SentiTextRank

Also the Url is as follows: https://github.com/Reasoning-NLG-Unito/SentiTextRank/tree/main.