B.M.S College of Engineering

P.O. Box No.: 1908 Bull Temple Road,

Bangalore-560 019

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING 2022-23

Course – Seminar Based on Summer/Winter Internship

Course Code – 20IS6SRITR

Seminar report on Internship

Submitted to – Dr. Roopa R

Submitted by – Nandan Hegde

B.M.S College of Engineering

P.O. Box No.: 1908 Bull Temple Road,

Bangalore-560 019

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING 2022-23

CERTIFICATE

Certified that the Seminar has been successfully presented at **B.M.S College Of Engineering** by Nandan Hegde bearing **USN 1BM20IS083** in partial fulfilment of the requirements for the VI Semester degree in **Bachelor of Engineering in Information Science & Engineering** of **Visvesvaraya Technological University, Belgaum** as a part of the course **Seminar Based on Summer/Winter Internship** during academic year 2022-2023.

Faculty Name – Dr.Roopa R

Designation – Assistant Professor

Department of ISE, BMSCE

TABLE OF CONTENTS

1. Problem Statement

The problem addressed in this literature survey is the need to comprehensively explore and understand sentiment analysis techniques in the context of social media. By reviewing and analyzing five selected research papers, the objective is to gain insights into the methodologies, findings, and advancements in sentiment analysis applied to social media data. The survey aims to provide a comprehensive overview of the topic, synthesizing the information from the papers to develop a deep understanding of the various approaches, challenges, and potential opportunities in sentiment analysis for social media. Through this study, the goal is to contribute to the existing knowledge and identify future directions for research and development in this rapidly evolving field.

2. Introduction

Sentiment analysis, also known as opinion mining, is a computational approach that aims to extract and analyze subjective information from textual data. With the widespread use of social media platforms, sentiment analysis has become a crucial tool for understanding public opinion, consumer behavior, and social trends. Social media platforms such as Twitter, Facebook, Instagram, and online review platforms have given individuals a platform to express their thoughts and opinions publicly. This abundance of user-generated content has opened up new opportunities for sentiment analysis, allowing researchers and organizations to tap into this vast amount of data to gain insights into public sentiment.

The field of sentiment analysis in social media has gained significant attention in recent years due to its potential applications in various domains. It has proven to be valuable in areas such as brand management, market research, political analysis, customer feedback analysis, and reputation management. By analyzing the sentiments expressed by users in social media posts, organizations can make informed decisions, understand public perception, and tailor their strategies accordingly.

The main objective of sentiment analysis in social media is to automatically determine the sentiment polarity (positive, negative, or neutral) of a given text. However, analyzing sentiment in social media poses unique challenges compared to traditional text analysis due to the informal nature of social media content, the use of slang, abbreviations, emoticons, and the presence of noisy and unstructured data.

To overcome these challenges, researchers and practitioners have developed various techniques and algorithms for sentiment analysis. These include rule-based approaches, lexicon-based methods, machine learning techniques, deep learning models, and hybrid approaches

In this literature survey, we will delve into the field of sentiment analysis in social media by studying five selected research papers. The survey aims to provide a comprehensive understanding of the methodologies, techniques, and advancements in sentiment analysis applied to social media data. By analyzing and synthesizing the findings from these papers, we aim to gain insights into the current state of the field, identify challenges and limitations, and explore potential future directions for research and development.

3. Pre-requisites

a. **Understanding of Sentiment Analysis**: Familiarize yourself with the basic concepts, techniques, and methodologies used in sentiment analysis.

b**. Text Processing and Natural Language Processing (NLP):** Develop a foundation in text processing techniques, including tokenization, stemming, stop word removal, and normalization.

c. **Machine Learning and Deep Learning**: Familiarize yourself with machine learning algorithms and techniques commonly used in sentiment analysis, such as Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Convolutional Neural Networks (CNN).

d. **Research Paper Analysis**: Develop skills in reading and analyzing research papers. Learn how to extract key information, identify research gaps, and understand the methodologies and findings presented in the papers.

e. **Evaluation Metrics:** Familiarize yourself with evaluation metrics commonly used to assess the performance of sentiment analysis models. These may include accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC).

f**. Ethical Considerations**: Recognize the ethical implications and challenges associated with sentiment analysis in social media, such as privacy concerns, biases in data and algorithms, and potential misuse of sentiment analysis results.

4. Literature Survey Details

Paper I

a. Title of Paper :

Sentiment Analysis on Social Media Data Using Intelligent Techniques

b. Author :

Kassinda Francisco Martins Panguila, Dr. Chandra J.

c. Problem Addressed :

The mixture of product and service reviews on Amazon makes it difficult for users to differentiate and make informed decisions based on sentiment analysis. The lack of clear bifurcation between product and service reviews, as well as the absence of sentiment analysis for specific product features, can mislead buyers. The paper proposes a system that addresses these challenges by performing classification of customer reviews, extracting sentiment for both product and service reviews, and visualizing the sentiment using charts.

d. Methodology used

The methodology of the study involves three main steps: data extraction, preprocessing, and feature representation, followed by the application of various machine learning algorithms for sentiment analysis. In the data extraction phase, relevant data, such as reviews and tweets, is collected for sentiment analysis.

Next, in the preprocessing stage, the collected data undergoes necessary transformations to prepare it for analysis. This may include steps like removing punctuation, tokenizing, and cleansing the text by eliminating irrelevant components like usernames, hashtags, and URLs.

After preprocessing, the data is represented using two approaches: sparse vector representation and dense vector representation. In the sparse vector representation, each word is assigned an integer index based on its frequency, while in the dense vector representation, a fixed-length vector is created for each survey using integer indices assigned to words.

The study applies seven different classifiers: Convolutional Neural Networks (CNN), Multi-layer Perceptron (MLP), Naïve Bayes, Maximum Entropy, Decision Tree, Random Forest, and Support Vector Machine (SVM). These classifiers are used to classify the sentiment of the data based on the chosen feature representation. The performance of the classifiers is evaluated by measuring their accuracy. To prevent overfitting, 10% of the training data is used for validation. The results indicate that dense vector representation performs better for sentiment analysis when implemented with Convolutional Neural Networks (CNN). The accuracies of the different classifiers are reported in Table III.

Overall, the study explores the effectiveness of various machine learning algorithms in sentiment analysis using different feature representations and provides insights into their performance.

e.Results

In this paper, the authors conducted experiments using various classifiers for sentiment analysis on social media data. The study employed a 10% validation dataset to prevent overfitting. The classifiers used include Naïve Bayes, Maximum Entropy, Decision Tree, Random Forest, SVM, and MLP, with the sparse vector representation of reviews and tweets. The results, as described in Table III, indicate that the sparse vector representation yielded better performance than frequency-based methods for sentiment analysis. Additionally, the dense vector representation was utilized specifically for implementing CNN, which achieved superior results compared to other classifiers.

The intelligent techniques approach proposed in this paper demonstrates the applicability of sentiment analysis on social media data. The experiments conducted on Twitter data and consumer affair website data, specifically related to Uber rides, highlight the effectiveness of neural network methods such as MLP and CNN. These methods outperformed other classifiers in general.

Paper II

a. Title of Paper :

Sentiment Analysis of Cyberbullying on Instagram User Comments

b. Author :

Muhammad Zidny , Novanda Alim Setya Nugraha

c. Problem Addressed :

The problem of cyberbullying on Instagram, which is a widely used photo-based social media platform. While Instagram allows users to share photos, videos, and engage with others through comments, the comment feature is often misused for cyberbullying. Despite the illegal nature of cyberbullying, Instagram lacks a built-in feature to detect and address such comments automatically. Therefore, this study aims to develop a system that can classify comments and detect cyberbullying elements. The proposed approach involves using the Naïve Bayes Classifier algorithm for comment classification, along with preprocessing and feature extraction using the TF-IDF method. The evaluation and testing of the system are performed using the K-Fold Cross Validation method, considering both stemmed and non-stemmed data. The experimental results demonstrate an accuracy of 83.53% for the best-performing model, which Includes the stemming process.

d. Methodology used

In the feature extraction stage, the focus is on determining the importance or weight of each term in the comments. The methodology used for weighting in this study is TF-IDF (Term Frequency - Inverse Document Frequency).

TF-IDF is a widely used technique in natural language processing that calculates the importance of a term in a document by considering its frequency in the document (term frequency) and its rarity in the entire document collection (inverse document frequency). This technique aims to capture the significance of a term within a specific document while also taking into account its overall occurrence across all documents.

The term frequency (TF) measures the number of times a term appears in a document, emphasizing the relevance of the term within that specific context. On the other hand, the inverse document frequency (IDF) measures the rarity of a term across the entire document collection, assigning higher weights to terms that are less common and potentially more informative.By combining the term frequency and inverse document frequency, TF-IDF provides a numerical weight for each term that reflects its importance in a given document. This weight can then be used as a feature representation for further analysis and classification tasks.

In summary, the feature extraction stage of this study employs the TF-IDF method to calculate the importance of terms in the comments, taking into account both their frequency in the document and their rarity in the entire document collection. This approach enables the system to capture the significance of terms and utilize them as features for the classification of cyberbullying comments..

e. Results

In conclusion, the research findings demonstrate that the Naïve Bayes Classifier algorithm is effective in classifying comments into cyberbullying and non-cyberbullying categories. Both the experiments with and without stemming at the preprocessing stage achieved the same best accuracy result of 84%. However, the use of stemming had a noticeable impact on the number of detected cyberbullying comments.

These results suggest that the proposed system utilizing the Naïve Bayes Classifier algorithm and TF-IDF feature extraction, along with preprocessing techniques such as folding case, cleansing, tokenizing, word replacing, stop words removal, and stemming, can effectively detect cyberbullying comments on Instagram. The findings also emphasize the importance of considering linguistic variations and word normalization techniques in preprocessing steps to enhance the accuracy of cyberbullying detection. Overall, this study contributes to addressing the issue of cyberbullying on Instagram by providing a systematic approach.

Paper III

a. Title of Paper :

Sentiment Analysis of Twitter Data: A Survey of Techniques

b. Author :

Vishal A. Kharde, S.S. Sonawane

c. Problem Addressed :

The paper acknowledges the vast amount of data generated on social networking sites like Twitter and the need to analyze and make sense of the unstructured and heterogeneous opinions present in the tweets. The paper provides a survey and comparative analysis of existing techniques for opinion mining, including machine learning and lexicon-based approaches. It also discusses the use of various machine learning algorithms such as Naive Bayes, Max Entropy, and Support Vector Machine for sentiment analysis on Twitter data streams.

The paper highlights the challenges and applications of sentiment analysis on Twitter and emphasizes the importance of sentiment analysis in understanding user opinions and preferences. It recognizes the significant role of social media in generating sentiment-rich data, such as tweets, status updates, blog posts, and comments. The paper mentions the influence of user-generated content on decision making and how sentiment analysis can help users evaluate products and services based on others' opinions.

d. Methodology used

The methodology used in the paper involves analyzing the Twitter dataset publicly available from Stanford University. The analysis is conducted on this labeled dataset using various feature extraction techniques. The framework includes a preprocessor applied to raw sentences, making them more suitable for understanding. Different machine learning techniques are then used to train the dataset with feature vectors, and semantic analysis provides a large set of synonyms and similarity to determine the polarity of the content.

The baseline algorithm used in the study is Naïve Bayes, without preprocessed data and using the unigram model. The accuracy obtained at different dataset sizes is presented in a table. The effect of stopwords is also investigated by removing them and running Naïve Bayes again. The results show a slight improvement in accuracy, indicating that stopwords have an impact on predictions. The removal of stopwords makes a noticeable difference in accuracy, which can be attributed to the common use of stopwords in short tweets.

Furthermore, the effect of bigrams is examined, where a combination of two words is used as a feature. Bigrams capture certain features in the data that unigrams fail to capture. The accuracy increases from the unigram model to the bigram model, demonstrating the effectiveness of bigrams in sentiment analysis. The results are presented in a table, showing the accuracy obtained at different dataset sizes for the Naïve Bayes algorithm with bigrams.

Overall, the methodology involves preprocessing the data, applying machine learning techniques, and exploring the effects of different features such as stopwords and bigrams on sentiment analysis accuracy.

e. Results

In conclusion, this paper provides a comprehensive survey and comparative analysis of existing techniques for opinion mining, encompassing machine learning and lexicon-based approaches. The research findings indicate that machine learning methods, such as SVM and naive Bayes, consistently achieve high accuracy and can be considered as baseline learning methods. On the other hand, lexicon-based methods show effectiveness in specific cases, particularly when there is limited availability of human-labeled data. The study also investigates the impact of various features on classifiers, revealing that cleaner data leads to more accurate results. Moreover, the utilization of the bigram model proves advantageous, as it enhances sentiment accuracy compared to other models.

Moving forward, future research should focus on integrating machine learning methods with opinion lexicon approaches to further enhance sentiment classification accuracy and adaptability across diverse domains and languages. This would enable the development of more robust and versatile sentiment analysis models. Additionally, exploring cross-domain and cross-lingual methods could facilitate sentiment analysis in varied contexts and languages, enabling a more comprehensive understanding of opinions and attitudes expressed in different cultural and linguistic settings. By addressing these areas of research, the field of sentiment analysis can advance its capabilities and contribute to a deeper understanding of user opinions and sentiments in online platforms.

Paper IV

a. Title of Paper :

Instagram Sentiment Analysis :Opinion Mining

b. Author :

Shweta Gangradea, Nirvishesh Shrivastavab, Jayesh Gangradea,

c. Problem Addressed :

Addresses the problem of sentiment classification on the Instagram platform. While conventional sentiment analysis approaches rely on polarity classification using sentiment lexicons, this study proposes a classification method based on Thayer's psychological model of emotions. The authors extract sentiment keywords for major sentiments by utilizing hashtags, which are prevalent in Instagram posts. Sentiments are determined by measuring the similarity between sentiment adjective candidates and the extracted sentiment keywords. This research contributes to the advancement of sentiment analysis techniques by incorporating psychologically defined sentiment categories and leveraging the unique characteristics of Instagram as a platform for expressing opinions and attitudes.

d. Methodology used

The paper focuses on conducting Instagram sentiment analysis by extracting words from image captions, including hashtags and emojis, in order to evaluate the effectiveness of these features for opinion mining. The goal is twofold: first, to extract all the words used in the user's Instagram captions, and second, to calculate the frequency of each distinct word used by the user across all their images.

The algorithm used in the Instagram analysis begins by taking the user's username as input. The script then retrieves and displays all the photos uploaded by the user, provided their profile is public. The data generated by the Python code needs to be refined and filtered to extract the meaningful information. Various filtering techniques are applied to remove unnecessary data and store the relevant content in Python arrays.

The refined data is then accessed for analysis. The paper provides an example where the most frequently used word, "swaranjali2k18," is highlighted in the upper corner and appears 23 times. Other words and their frequencies are also identified. This approach allows organizations to gain insights into users' thought processes, interests, and opinions about various topics such as products, politics, and preferences. By analyzing Instagram captions, organizations can anticipate users' demands and preferences even before the users themselves are aware of them.

Overall, the methodology involves extracting words from image captions, refining and filtering the data, and analyzing word frequencies to gain valuable insights into user sentiment and preferences on Instagram. This approach can be beneficial for businesses and organizations seeking to understand their audience and tailor their strategies accordingly.

e. Results

sentiment analysis on social-photo sharing platforms like Instagram has gained significant attention in recent years. Companies and media organizations are increasingly interested in mining Instagram data to understand people's sentiments towards their products and services. The application of sentiment detection extends to various domains, including review classification and real-time analysis. It is evident from the research that different classification models and features have distinct distributions, and their performance varies across domains. Future work should focus on improving performance measures and addressing challenges such as multilingual analysis, handling negation expressions, and generating opinion summaries based on product features. There is still much research needed to overcome these challenges and further advance the field of sentiment analysis.

Paper V

a. Title of Paper :

Twitter sentiment analysis using hybrid gated attention recurrent network

b. Author :

Nikhat Parveen, Prasun Chakrabarti, Bui Thanh Hung and Amjan Shaik

c. Problem Addressed :

The paper addresses several problems related to sentiment analysis using deep learning techniques on Twitter data. To overcome these challenges, the proposed architecture in the paper combines recurrent neural networks (RNN) and attention mechanisms for sentiment analysis on Twitter data. The architecture leverages the LTF-MICF model to extract sentiment-based features from the pre-processed dataset. These features help improve the efficiency and effectiveness of the classifier. To address the dimensionality issue caused by a large number of extracted features, the paper introduces the hybrid mutation-based white shark optimizer (HMWSO) algorithm for feature selection, reducing the dimension occupied by the features.

By combining RNN, attention mechanisms, LTF-MICF feature extraction, and HMWSO feature selection, the proposed architecture aims to overcome the shortcomings of existing algorithms, including long processing times, complexity, and accuracy limitations. The goal is to develop an efficient and effective approach for sentiment analysis on Twitter data.

d. Methodology used

The proposed methodology for sentiment classification of Twitter tweets involves the use of a DL technique called Gated Attention Recurrent Network (GARN). The initial step is to collect the Twitter dataset, specifically the Sentiment140 dataset, which consists of sentiment-labeled tweets accessible to the public. Once the dataset is collected, the pre-processing stage is performed. This stage involves various text cleaning operations such as tokenization, removing stopwords, stemming, correcting slang and acronyms, removing numbers, punctuation, symbols, converting uppercase to lowercase, and removing hashtags, user mentions, characters, and URLs.

After pre-processing, the pre-processed dataset serves as the input for the next step. In the feature extraction phase, the Log Term Frequency-based Modified Inverse Class Frequency (LTF-MICF) technique is applied. This technique assigns a term weight to each term in the dataset based on its term frequency. The LTF-MICF extraction technique helps in capturing sentiment-based features from the pre-processed data.

To select the optimal term weights and reduce the dimensionality of the features, the Hybrid Mutation-based White Shark Optimizer (HMWSO) is utilized. The HMWSO algorithm performs feature selection by iteratively mutating and optimizing the feature weights, leading to an optimal subset of features that contribute most to sentiment classification.

Finally, the output of HMWSO, which consists of the selected term weights, is fed into the GARN model for sentiment classification. GARN is a DL model that combines recurrent neural networks (RNN) and attention mechanisms. It takes the selected features as input and performs sentiment classification, categorizing the tweets into three different classes: positive, negative, and neutral.

Overall, the proposed methodology involves a multi-step process, starting from data collection and pre-processing, followed by feature extraction using LTF-MICF, feature selection using HMWSO, and sentiment classification using GARN. The combination of these techniques aims to improve the efficiency and accuracy of sentiment analysis for Twitter tweets.

e. Results

GARN is preferred in this research to find the various opinions of Twitter online platform users. The implementation was carried out by utilizing the Sentiment 140 dataset.The performance of the leading GARN classifier is compared with other DL models Bi-GRU, Bi-LSTM, RNN and CNN for four performance metrics: accuracy, precision, f-measure and recall centred with four-term weighting schemes LTF-MICF, TF-DFS, TF-IDF, TF and W2V. The evaluation shows that the leading GARN DL technique reached the target level for Twitter sentiment classification. Additionally, while applying the suggested term weighting scheme-based feature extraction technique LTF-MICF with the leading GARN classifier gained an efficient result for tweet feature extraction. With the Twitter dataset, the GARN accuracy on applying LTF-MICF is 97.86%. The accuracy value attained by the proposed classifier is the highest of all the existing classifiers. Finally, the suggested GARN classifier is regarded as an effective DL classifier for Twitter sentiment analysis and other sentiment analysis applications. Further, this method is analysed using the small dataset, therefore in future large data with challenging images will be used to analyse the performance of present architecture.

Paper VI

a. Title of Paper :

TOURISM PRODUCTS USING AN ASPECT-BASED OPINION MINING APPROACH

b. Author :

Edison Marrese-Taylor a, Juan D. Velásquez a, Felipe Bravo-Marquez b, Yutaka Matsuo c

c. Problem Addressed :

The problem discussed in this paper is the need to analyze customer preferences in the tourism domain using aspect-based opinion mining techniques. With the growth of Web 2.0 and social networks, individuals and enterprises are increasingly relying on user-generated content to make better decisions. In the case of the tourism industry, understanding customer preferences is crucial for providing better services and experiences.

The authors propose an extension of Liu's aspect-based opinion mining methodology to apply it specifically to the tourism domain. Traditional approaches in opinion mining mainly focus on physical product reviews, where users directly mention product features they liked or disliked. However, for tourism products like restaurants or hotels, users tend to provide more detailed and storytelling reviews that may include aspects beyond the product itself, such as the ambience, setting, or people involved.

The authors identify several challenges in analyzing tourism product reviews, including longer and more complex sentences, multiple mentions of aspects, the presence of non-aspect objects, and sentences without opinions. Existing approaches do not adequately address these issues, so the authors propose new rules and techniques to handle these complexities.

d. Methodology used

1. Aspect Identification: The first step is to identify and extract important topics (aspects) from the text. The proposed technique uses part-of-speech tagging, syntax tree parsing, and frequent itemset mining to extract frequent nouns and noun phrases. Linguistic rules are then applied to filter and eliminate redundant aspects.

2. Sentiment Prediction: The second step involves determining the sentiment orientation for each aspect. The approach described relies on a sentiment word dictionary and linguistic rules. Positive and negative opinion words are matched in the text, and special rules handle cases of negation and intensifiers like "too." An aggregation score function is used to determine the overall sentiment orientation of an aspect in a sentence.

3. Summary Generation: The final step is to generate a summary that presents the processed results in a simple manner. The proposed approach suggests using aspect-based opinion summaries, which are bar charts showing the number of positive and negative opinions for each aspect. These bar charts can be used to compare products and display the percentages of positive and negative reviews.

The extension builds upon the previous steps and addresses specific challenges in tourism product reviews. It introduces a model to extract opinion tuples from opinionated documents. The model considers sentences, aspect expressions, and entity expressions to determine sentiment orientation for each pair. Additionally, the extension proposes rules for handling compound aspect expressions and multiple occurrences of aspects in a sentence. Furthermore, the extension suggests measuring the importance of each aspect by considering both the number of positive and negative opinions. The assumption is that aspects with a higher number of opinions in both orientations are more important. Standard deviation is used to quantify the dispersion of positive and negative opinion scores, allowing for ranking aspects based on importance.

Overall, the proposed extension aims to enhance aspect-based sentiment analysis by addressing the challenges specific to tourism product reviews and providing a measure of aspect importance for summarization purposes.

e. Results

- Explicit Aspect Extraction: The algorithm achieved a precision of 33% for hotels and 42% for restaurants, and recall of 29% for hotels and 37% for restaurants.

- Subjectivity Classification: The algorithm achieved an average precision of 80% and recall of 91%.

- Sentiment Classification: The algorithm achieved an average precision of 90% and recall of 93%.

The results indicated that the performance of the aspect extraction task was poor, with the algorithm only capturing a small percentage of explicit expressions. Many of the extracted expressions did not correspond to real aspect expressions. However, sentiment classification showed good results, although it was evaluated only for the extracted aspect expressions. The presence of non-opinionated sentences in tourism product reviews contributed to a decrease in precision for subjectivity classification.

In summary, the experiment showed that while sentiment classification performed well, the aspect extraction task had poor results, indicating difficulties in accurately identifying explicit aspect expressions in the tourism domain.

Paper VII

a. Title of Paper :

Sentiment Analysis Of Text Using Machine Learning Models

b. Author :

P Ancy Grana

c. Problem Addressed :

The aim is to develop and implement sentiment analysis models to extract and analyze sentiment from textual data. The papers focus on various aspects of sentiment analysis, including data collection, data preprocessing, feature extraction, and machine learning classification techniques. The objective is to accurately determine the polarity (positive, negative, or neutral) of the sentiments expressed in the text and assess the performance of different sentiment analysis approaches.

The research paper also address the challenges associated with sentiment analysis, such as handling subjective opinions, contextual understanding, and sarcasm detection. They propose different methodologies, including hybrid approaches that combine sentiment analyzers and machine learning classifiers, as well as the application of both traditional machine learning algorithms (Naïve Bayes Classifier, Support Vector Machine, Decision Tree) and deep learning models. The papers aim to evaluate and compare the performance of these methods to determine their effectiveness in sentiment analysis tasks.

c. Methodology used :

The methodology used in the mentioned research paper for sentiment analysis involves several steps, namely data collection, data preprocessing, feature extraction, and sentiment classification. These steps aim to gather relevant data, make it more machine-readable, extract meaningful features, and classify the sentiment of the text using rule-based and automatic approaches.

1. Data Collection:

The researchers gather relevant tweets from the Twitter API related to the specific subject of interest. The collected dataset is crucial for the efficiency of the sentiment analysis model. The dataset is divided into training and testing sets, which play a significant role in determining the model's performance.

2. Data Preprocessing:

Data preprocessing is a crucial step to enhance the efficiency of subsequent analysis. It involves various steps to make the data more machine-readable and reduce ambiguity in feature extraction. Some of the preprocessing steps mentioned in the paper include removing retweets, non-English words, stop words, special characters, and digits. Tokenization, POS tagging, stemming, spelling correction, and expansion of slangs and abbreviations are also performed. Additionally, a dictionary is created to remove unwanted words and punctuation marks.

3. Feature Extraction:

Feature extraction is performed to select useful words from the tweets, which will serve as features for sentiment analysis. The paper mentions different techniques for feature extraction, including unigram and n-gram features. External lexicons are used, which consist of predefined positive or negative sentiment words. Frequency analysis is conducted to identify the most frequently used features in the dataset.

4.Sentiment Classification:

The sentiment classification step involves determining the sentiment polarity of the text. The paper discusses two approaches: rule-based and automatic sentiment analysis.

- Rule-Based Approach: The rule-based algorithm utilizes a set of predefined rules to classify sentiment. It involves processes such as stemming, tokenization, POS tagging, parsing, and lexicon analysis. The algorithm compares words in the text with positive and negative word lists to determine the prevalent sentiment.

- Automatic Sentiment Analysis Approach: The automatic approach utilizes supervised machine learning algorithms for classification and unsupervised machine learning approaches for data exploration. The paper mentions specific algorithms used for sentiment analysis, including Naïve Bayes, linear regression, and support vector machines (SVM). These algorithms are trained on the extracted features to predict sentiment polarity.

These methodologies aim to classify reviews into service reviews, feature reviews, and product reviews, while also extracting sentiment and visualizing the results. The algorithm provides a systematic approach to handle various aspects of sentiment analysis on Amazon customer reviews, offering insights into both the sentiment expressed and the classification of different review types.

e. Results :

Sentiment analysis is an invaluable technology for businesses due to its ability to accurately extract information from customer feedback and public opinion in an unbiased manner. When implemented effectively, it can provide significant value to systems, applications, and web services. By leveraging machine learning, sentiment analysis enables businesses to analyze public opinion, improve customer support, and streamline operations with efficiency. This not only saves time but also reduces costs. Furthermore, the insights derived from sentiment analysis yield actionable information that aids in making well-informed decisions.

In this research paper, we employed various machine learning algorithms to perform sentiment analysis. We implemented and compared two distinct approaches: a rule-based approach utilizing NLP algorithms and an automatic sentiment analysis approach using ML algorithms. These methods facilitated the identification of sentiment using advanced ML techniques. By incorporating sentiment analysis into their operations, businesses can gain a deeper understanding of public sentiment, enhance customer support services, and make data-driven decisions.

Paper VIII

a. Title of Paper :

Graph-Based Conversation Analysis in Social Media

b. Author :

Marco Brambilla 1,∗ , Alireza Javadian Sabet 2 , Kalyani Kharmale 3 and Amin Endah Sulistiawati

c. Problem Addressed :

It is crucial to understand the communication behavior between SM users. For instance, when users express their idea through comment sessions on an SM post, conversations are created at least between the author and the engaged users. These formed conversations among SM users are the core of virtual communication that deputizes closely to real communication. Since most studies on SNs are focused on a user-to-user relationship, they sometimes miss the crucial information from the conversations, i.e., the user-generated content (UGC). These UGC are fundamental to conceive online communication behavior. Considering a large dataset from SM platforms with its complex structure, the research questions that lead to this work are as follow:

1. How to build a proper graph for describing the conversational aspect of online SM?

2. How to reconstruct conversations from comments belong to an SM post/update that does not follow reply feature?

3. How to assign an appropriate category label to an SM comment that represents the author’s intention?

4. How to uncover micro topics that are discussed under one main topic.

5. How are the topics, stance, and sentiments propagate on the discussion forums?

6. What frequent patterns can be found in conversation graphs of online SM?

d. Methodology used:

The methodology proposed in this work consists of three main stages: Web Scraping, Text Processing, and Network Design. Here is a breakdown of each stage:

1. Data Collection (Web Scraping): Design a model to collect data from social media platforms. Use automated programs to scrape web pages and parse the data into JSON format. Store the collected data in a database with JSON-like document schemas.

2. Data Cleaning and Preprocessing (Text Processing): Remove records with missing values. Apply text cleaning to remove unwanted characters and typos.Perform stemming to produce a bag-of-words representation of the text.Compute TF/IDF scores to obtain a word/document weight matrix.

3. Text Classification Design: Define the desired categories (classes) for text classification with the help of domain experts. Use keyword-based classification to assign labels to comments based on the presence of specific keywords. Apply machine learning classification algorithms (e.g., Naïve Bayes, Support Vector Machine) to classify remaining unclassified comments. Involve human validation in the process to improve accuracy.

4. Sentiment Analysis: Perform sentiment analysis to detect positive, negative, or neutral sentiment in sentences. Use the TextBlob package in Python for sentiment analysis based on polarity scores.

5. Topic Modeling: Utilize Latent Dirichlet Allocation (LDA) to detect micro topics within a main topic. Use LDA to cluster words into topics and infer the distribution of topics in the text.

6. Stance Detection: Identify the author's view or stance about a topic or target. Use a supervised machine learning approach to classify comments as in favor, against, or none. Train models on algorithms like Support Vector Machine, Random Forest Classifier, and Neural Networks MLP Classifier.

7. Network and Conversation Graph Design: Design a directed multigraph to represent relationships between social media content (posts, users, comments, etc.). Define nodes for posts, users, comments, hashtags, locations, etc., and connect them with appropriate edges. Generate the graph representation of the relationships among social media content. Retrieve and store the generated graph for further analysis.

The proposed methodology covers tasks such as data collection, data cleaning, text processing, text classification, sentiment analysis, topic modeling, stance detection, and network graph construction. Each stage contributes to the overall goal of understanding communication behavior and dynamics on social media platforms.

e. Results:

The study aimed to compare communication behavior in social media (SM) discussions with real-life conversations. A graph-based framework was proposed for analyzing online conversations. Intent analysis using keyword-based classification was employed for social media comments. The study focused on Instagram posts related to the "YourExpo2015" challenge.

Comments were classified into nine categories based on predefined keywords. Naïve Bayes and SVM algorithms were used to process uncategorized comments, with human-in-the-loop feedback for refinement. The overall accuracy of the classification approach was 98%. A directed multigraph represented the collected SM dataset, with nodes for posts, comments, authors, locations, and hashtags. The graph had over 450K nodes and 1.4M edges.

An experiment on COVID vaccine-related discussions on Reddit was conducted, employing topic modeling, sentiment analysis, and stance detection. Discussions received the most engagement within the first 6 hours. Negative topics introduced early could influence the discussion thread. Most statements lacked a specific stance, with queries and questions prevailing. Despite the impact of COVID, discussions contained positive content.

Future research will explore intent analysis in more depth, enhance graph analysis with community detection, and investigate feature selection methods. Emoji and emoticon symbols may be considered, and conversation agents based on learned patterns could facilitate customer relationship management and behavioral changes in users.

Paper IX

a. Title of Paper :

Graph partitioning and visualization in graph mining: a survey

b. Author :

Swati A. Bhavsar & Varsha H. Patil & Aboli H. Patil

c. Problem Addressed :

This paper provides a comprehensive survey of published work in the field of graph mining. Graph mining involves obtaining sub-graphs from a larger graph and has applications in various domains such as social network analysis, computer network design, bioinformatics, and image processing. The paper categorizes the different graph mining algorithms and provides an overview of their basic concepts and contributions by various authors. It also discusses common issues in graph mining, including clustering, partitioning, and graph visualization. Additionally, the paper mentions standard datasets available for graph mining. The problem addressed in the paper is the need for a comprehensive understanding of graph mining algorithms and their applications.

d. Methodology used:

Graph mining techniques can be categorized into three main types: graph clustering, graph classification, and sub-graph mining.

Graph clustering: In graph clustering, the vertices of a graph are grouped together to form clusters based on the edge structure of the graph and similarities of graph structures. The goal is to create clusters in such a way that there are more inter-cluster edges and fewer intra-cluster edges. Graph clustering is typically performed using unsupervised learning techniques since the classes are not known in advance.

Graph classification: In graph classification, the goal is to classify an entire graph into separate classes or categories. This classification is based on supervised or semi-supervised learning techniques, where the classes of the data are known in advance. The input graph is assigned a class label based on its characteristics or features.

Sub-graph mining: Sub-graph mining involves finding frequent sub-graphs, which are graphs whose vertices and edges are subsets of another graph. The frequent sub-graph mining problem aims to identify the set of sub-graphs that occur in at least a given threshold of the input example graphs. This technique is often used in pattern mining and can be applied to various domains.

Navigating large graphs is challenging due to limited space on small displays. Challenges include:

1. Limited overview: It's hard to see the entire structure of a large graph on a small display, potentially missing patterns or relationships.

2. Zooming and panning: Zooming out loses details, while zooming in may cause overlaps or unreadable elements.

3. Node identification: Locating specific nodes is difficult when there are many densely packed nodes or unclear labels.

4. Link exploration: Tracing connections between nodes becomes daunting, requiring extensive scrolling or searching.

Solutions include:

- Optimized zooming and panning algorithms.

- Filtering and grouping nodes based on criteria.

- Highlighting selected nodes or links.

- Search and navigation aids like overview maps.

These techniques enhance readability and enable efficient exploration of large graphs on small displays.

e. Results:

This survey paper provides an overview of the advancements in graph routing, analysis, and visualization over the past few decades. It examines various literature sources such as journals, transaction papers, patents, reviews, and surveys to understand the research landscape and identify research gaps. The paper focuses on graph mining, proposing a comprehensive taxonomy and discussing the techniques, advantages, and disadvantages of different graph mining approaches. It also addresses key challenges in graph partitioning, graph classification, and graph visualization, highlighting commonly used datasets for different applications. Graph partitioning is identified as a crucial and difficult process, aiming to accelerate design, maintain system functionality, simplify tasks, and minimize interactions between blocks. The accuracy of partitioning directly affects its quality and the preservation of connectivity through edges.

Paper X

a. Title of Paper :

A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text

b. Author :

Ning Liu and Jianhua Zhao

c. Problem Addressed :

The problem discussed in this paper is cross-domain text sentiment classification. With the increasing amount of subjective text data with emotions on social media platforms, there is a need to extract emotional information from these texts for various applications such as product recommendation, customer management, and news analysis. However, sentiment analysis is domain-dependent, and models trained on one domain may not perform well on another domain. Building labeled datasets for each domain is expensive and time-consuming.

d. Methodology used

The methodology described in the paper involves cross-domain text sentiment analysis using a combination of BERT model, convolutional model (CNN), and adversarial model. The goal is to train a sentiment classifier on a labeled source domain dataset and then use it to classify sentiment on an unlabeled target domain dataset.

The algorithmic framework consists of several steps. First, the input data, including sentence and aspect word representations from both domains, undergoes feature extraction using BERT and CNN with shared weights. BERT extracts the semantic information of the sentences, and CNN further extracts key local features. Dimensionality reduction is applied to the features with high semantic information. The output features of CNN serve as inputs to adversarial classifiers and sentiment classifiers. The domain adversarial classifier aims to confuse the domain information, while the sentiment classifier focuses on aspect-level sentiment classification.

To improve the performance of the CNN, the structure is modified by adding a gated activation unit. This modification assigns higher weights to emotional words that are closer in aspect information, thereby enhancing aspect-level sentiment analysis.

During model training, the sentiment-labeled data from the source domain is used to train the sentiment classifier, while the unlabeled data from the target domain is combined with the features extracted from the source domain and used as input for the domain adversarial classifier. The errors from both classifiers are backpropagated to update and optimize the model parameters.

The implementation process involves using BERT for sentence-level sentiment classification and treating aspect-level sentiment classification as a sentence pair classification task. The special tokens "[CLS]" and "[SEP]" are used for text representation, and the aspect-level sentence is represented by the output vector corresponding to "[CLS]". A classifier consisting of convolutional layers and Softmax layers is applied for sentiment classification.

The text convolution process involves convolving the output of BERT using a modified CNN, which extracts local text features and reduces the dimension of shared emotional features. Convolution kernels of different sizes are used, and the results are merged and subjected to max pooling.

The sentiment classifier is responsible for the source domain dataset and performs sentiment classification using the features obtained after convolution pooling. Softmax is applied to predict sentiment classification.

Domain confrontation aims to generalize the features from the source domain to the target domain, making it difficult for the classifier to differentiate between the domains. A domain classifier is trained using logistic regression as a domain adversarial structure. The features extracted from different domains are input together into the domain classifier, which cannot distinguish the domain of the features.

The objective function of the model combines the loss functions of the sentiment classifier and the domain adversarial training. The sentiment classifier loss function uses labeled data from the source domain, while the domain adversarial loss function incorporates a mixture of source and target domain labels. The parameters of the model are updated using the backpropagation algorithm.

In summary, the methodology involves combining BERT, CNN, and adversarial models to perform cross-domain sentiment analysis. The approach leverages feature extraction, convolutional processing, sentiment classification, and domain confrontation to achieve aspect-level sentiment classification on target domain data using a sentiment classifier trained on the source domain data.

e. Results

The experiment conducted in the study involved the use of the Amazon product review dataset for cross-domain aspect-level sentiment analysis. The dataset consisted of reviews from five different domains: Books, DVD disks, Electronics, Kitchen appliances, and Videos. The data was manually annotated to create a balanced and suitable dataset for the task. The experiment compared the proposed method with and without BERT preprocessing and with and without a gated activation unit.

The results showed that the aspect-level cross-domain sentiment analysis with BERT preprocessing achieved better accuracy and F1 scores compared to the method without BERT preprocessing. The inclusion of the gated activation unit also improved the results compared to sentence-level sentiment analysis without gating.

Furthermore, the proposed method was compared with other existing methods such as SCL-ML, ITIAD, and CGRU. The results demonstrated that the proposed method outperformed these methods, achieving higher accuracy and F1 scores. The average improvement in accuracy compared to these methods was 6.4%, 4.1%, and 2.0% respectively.

Overall, the experiment validated the effectiveness of the proposed method for fine-grained cross-domain sentiment analysis. It showcased the advantages of utilizing BERT preprocessing and a gated activation unit, resulting in improved classification accuracy and performance across various domains.

5. Future Scope of Sentiment Analysis :

Sentiment analysis, also known as opinion mining, has made significant strides in analyzing and understanding the sentiment expressed in social media data. However, there are several areas where further advancements and research can contribute to the future development and improvement of sentiment analysis in social media.

1. Fine-grained sentiment analysis: Currently, most sentiment analysis techniques focus on classifying sentiments into broad categories such as positive, negative, or neutral. Future research can explore more fine-grained sentiment analysis, where sentiments are classified into more nuanced categories or even quantified on a continuous scale. This would provide more detailed insights into the sentiment expressed in social media posts.

2. Contextual understanding: Context plays a crucial role in understanding sentiment accurately. Sentiment analysis models should be capable of capturing and incorporating contextual information such as sarcasm, irony, or cultural references that can greatly affect the sentiment expressed in social media posts. Developing techniques to better handle contextual understanding will improve the accuracy and reliability of sentiment analysis.

3. Multimodal sentiment analysis: Social media content consists not only of text but also images, videos, and audio. Incorporating multimodal sentiment analysis techniques that can analyze sentiments expressed through different modalities will provide a more comprehensive understanding of sentiment in social media data. This involves exploring methods that can effectively extract sentiment from visual and auditory cues in addition to textual data.

4. Domain adaptation and transfer learning: Sentiment analysis models trained on one domain often struggle to perform well in another domain due to variations in language, vocabulary, and sentiment expressions. Future research should focus on domain adaptation and transfer learning

techniques that enable sentiment analysis models to generalize well across different domains. This would enhance the versatility and applicability of sentiment analysis in various social media contexts.

5. Real-time analysis: With the rapid growth of social media and the continuous stream of user-generated content, real-time sentiment analysis becomes crucial. Developing efficient and scalable algorithms that can process and analyze social media data in real-time will enable businesses and organizations to respond promptly to emerging trends and sentiments, making sentiment analysis more actionable and valuable.

6. Ethical considerations and bias mitigation: Sentiment analysis models need to address ethical concerns such as bias in data, fairness, and privacy. Future research should focus on developing techniques to identify and mitigate biases present in social media data and sentiment analysis models, ensuring fair and unbiased sentiment analysis results.

In conclusion, the future of sentiment analysis in social media holds tremendous potential for advancements in fine-grained analysis, contextual understanding, multimodal analysis, domain adaptation, real-time analysis, and addressing ethical considerations. Continued research and innovation in these areas will further enhance the accuracy, versatility, and usefulness of sentiment analysis in understanding and harnessing the power of social media data.

6. Conclusion

In conclusion, sentiment analysis in social media has emerged as a valuable tool for understanding and analyzing the sentiment expressed by users across various platforms. The research papers reviewed in this study have provided insights into the methodologies, techniques, and algorithms used in sentiment analysis, as well as their effectiveness in classifying sentiments and detecting specific aspects such as cyberbullying.

The papers highlight the importance of preprocessing techniques such as data collection, labeling, cleansing, tokenizing, and stemming to prepare the data for sentiment analysis. Additionally, feature extraction methods like TF-IDF have been shown to be effective in representing the importance of terms in documents, aiding in sentiment classification.

Various machine learning algorithms, including Naïve Bayes, Maximum Entropy, Decision Tree, Random Forest, SVM, Multi-layer Perceptron (MLP), and Convolutional Neural Networks (CNN), have been explored for sentiment analysis in social media. These algorithms have demonstrated promising results in accurately classifying sentiments and achieving high accuracy levels. Moreover, domain adaptation, real-time analysis, and ethical considerations are crucial aspects that need to be addressed in the future. Adapting sentiment analysis models to different domains, developing real-time analysis capabilities, and mitigating biases and ensuring fairness in sentiment analysis results are key challenges to overcome.

7. References

[1] Kassinda Francisco, Dr. Chandra J. “Sentiment Analysis on Social Media Data Using Intelligent Techniques.” ,International Journal of Engineering Research and Technology. ISSN 0974-3154 (2019).

[2] Muhammad Zidny Naf'an, Novanda Alim Setya Nugraha, ” Sentiment Analysis of Cyberbullying on Instagram User Comments” (2019).

[3] Vishal A. Kharde, S.S. Sonawane. ” Sentiment Analysis of Twitter Data: A Survey of Techniques.” International Journal of Computer Applications (0975 – 8887) (April 2016)

[4] Shweta Gangrade, Nirvishesh Shrivastava, Jayesh Gangrade, ”Instagram Sentiment Analysis: Opinion Mining.” (2019)

[5] Nikhat Parveen, Prasun Chakrabarti, Bui Hung and Amjan, ” Twitter sentiment analysis using hybrid gated attention recurrent network.” Journal of Big Data(2023)

[6] Edison Marrese-Taylor a, Juan D. Velásquez a, Felipe Bravo-Marquez b, Yutaka Matsuo c, TOURISM PRODUCTS USING AN ASPECT-BASED OPINION MINING APPROACH.

[7] P Ancy Grana*1, Jawaharlala Nehru Engineering College, MGM University, India, SENTIMENT ANALYSIS OF TEXT USING MACHINE LEARNING MODELS

[8] Marco Brambilla 1,∗ , Alireza Javadian Sabet 2 , Kalyani Kharmale 3 and Amin Endah Sulistiawati "Graph-Based Conversation Analysis in Social Media"

[9] Swati A. Bhavsar & Varsha H. Patil & Aboli H. Patil "Graph partitioning and visualization in graph mining: a survey"

[10] Ning Liu and Jianhua Zhao, A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text.