UNIVERSITY OF BERGEN

Department of Information Science and Media Studies

MASTERS THESIS

Developing a conceptual framework for sentiment analysis using LLMs

Author: Stian Vedvik Århus

Supervisor: Fazle Rabbi

June 1, 2023

Abstract

This study aims to develop a conceptual framework for sentiment analysis using Large Language Models (LLMs) in the context of social media data, with a focus on the importance of context in improving the reliability of sentiment analysis. To achieve this goal, the study employs a design science research methodology, where the proposed framework is tested and refined through a series of iterations. Specifically, contextual features such as user demographics, temporal factors, and domain-specific features are tested to provide more trustworthy sentiment analysis of social media data. The study presents a review of existing literature on sentiment analysis, and LLMs, and proposes a novel framework that integrates these approaches. Furthermore, this research addresses the challenges of the hallucination problem in sentiment analysis and emphasizes the importance of providing explanations and transparency in AI systems. It also recognizes the availability of AI tools to individuals without deep knowledge of how they work and highlights the need for responsible and ethical deployment of AI technologies.

The proposed framework is evaluated by comparing the sentiment analysis using LLMs with and without the incorporation of contextual factors, demonstrating the effectiveness of the context-aware approach in improving the reliability of sentiment analysis in the domain. The evaluation is conducted using a human survey, where participants with different backgrounds are asked to assess the sentiment of social media texts. The results of the survey are then compared with the sentiment analysis results generated by the proposed framework.

The framework has the potential to improve the use of sentiment analysis in various domains, such as political analysis and journalistic work. This study emphasizes the importance of design science research in the development of effective and practical solutions for sentiment analysis in social media data and highlights the need for further research to improve the scalability and generalizability of the proposed framework.

Acknowledgment

I want to express my gratitude to my supervisor, Fazle Rabbi for his significant contribution to my research work. The guidance and support have been valuable, and I could not have completed this thesis without your input. Your willingness to share your expertise, provide constructive feedback, and offer encouragement has been appreciated. Thank you for your dedication and for challenging me to improve. I am honored to have had you as my supervisor.

S.V.Å.

Contents

List of Figures

List of Tables

List of abbreviations

LLMs Large Language Models

USD United States dollar

NLP Natural language processing

SVM Support Vector Machines

SSWE Sentiment-specific word embedding

BERT Bidirectional Encoder Representations from Transformers

RNNs Recurrent neural networks

LSTM Long short-term memory networks

GRU Gated Recurrent Unit

GPT Generative Pre-trained Transformer

AI Artificial Intelligence

SMS Short Message Service

GLUE General Language Understanding Evaluation benchmark

MultiNLI Multi-genre Natural Language Inference

SQuAD The Stanford Question Answering Dataset

RLHF Reinforcement Learning from Human Feedback

RM Reward Model

PPO Proximal Policy Optimization

IT Information Technology

DSR Design Science Research

PGMOL Professional Game Match Officials Limited

Chapter 1

Introduction

The advent of social media platforms has impacted the way individuals express their thoughts, opinions, and emotions in the digital realm. According to a report by ResearchandMarkets, the global sentiment analysis market is projected to reach USD 10.1 Billion by 2030 [4]. The increasing demand for social media monitoring and the growing use of sentiment analysis in customer experience management and branding as some of the key factors driving the market growth. Twitter has emerged as a platform where you can yourself about different topics happening in the world. It is a source of real-time, human-generated data, making it a rich and valuable resource for, among other things, sentiment analysis. Sentiment analysis, also known as opinion mining, is a computational approach that aims to extract and analyze sentiments, attitudes, and emotions from textual data. It has gained significant attention in recent years due to its potential to uncover valuable insights into public opinions, political sentiments, and social trends all around the world.

Twitter's unique characteristics present both opportunities and challenges for sentiment analysis. The platform's microblogging format imposes a strict limit on the length of individual tweets, often resulting in fragmented and context-dependent expressions. Furthermore, the informal language, use of slang, abbreviations, and the prevalence of hashtags and emoticons pose additional complexities for accurate sentiment analysis. These challenges call for a structured approach and a robust conceptual framework that can effectively handle the peculiarities of sentiment analysis on Twitter. This content provides a valuable source of data that can provide insights into people's opinions and attitudes about various topics but has the potential to be biased displayed [2]. With sentimental analysis, we can identify the emotional state of a user, such as whether a tweet expresses a positive, negative, or neutral sentiment. In this thesis, an angle of how for example journalists can gain insights and easier collect people's opinions are in mind. Because of the complexity of social media content, in mind that it is often short-form, with limited context, and includes slang, misspellings, and

other forms of non-standard language. As a result, traditional sentiment analysis methods, such as rule-based or lexicon-based approaches, are often insufficient and may produce inaccurate results [8].

Sentiment analysis is not a simple process, and when models are investigating huge amounts of data automatically, it can be hard to know if you can trust the results the models give you. The complexity of human language, the use of sarcasm and irony, and the cultural nuances of language can all make it difficult to accurately determine the emotional tone of a piece of text.

The purpose of a contextual-based analysis is that it takes into account the surrounding words and phrases that help to determine the intended emotional tone of a piece of text. This approach is often more accurate than non-contextual analysis, as it allows for a more nuanced understanding of the text. However, contextual-based analysis can be more computationally expensive and requires more processing power than non-contextual analysis.

Non-contextual sentiment analysis has historically looked at individual words and phrases to determine the emotional tone. This approach is often simpler and more computationally efficient than contextual-based analysis, but it can also be less trustworthy for the recipient, as it does not take into account the nuances of the surrounding text.

This research paper aims to develop a conceptual framework for sentiment analysis using human-generated data on Twitter. The proposed framework will through design science research look at the impact context can have on the sentiment analysis via prompts. By integrating existing methodologies, techniques, and insights from the field of natural language processing and machine learning, this framework will address the specific requirements and challenges that can be associated with sentiment analysis on human-generated data from Twitter. The conceptual framework will encompass components involved in sentiment analysis on Twitter, including data collection, state-of-the-art technologies and comparison between results from sentiment analysis with both non-contextual and contextual information provided. It will emphasize the importance of considering the unique characteristics of Twitter data, such as noise, sarcasm, ambiguity, and the contextual dependencies that influence sentiment expressions. By incorporating the idea of these aspects into the framework, it aims to improve the quality and awareness of sentiment analysis results. The outcome of this research will be a practical and structured approach for sentiment analysis of human-generated data on Twitter. It will empower researchers and practitioners to gain deeper insights into public sentiment, brand perception, and social trends by effectively harnessing the vast amount of user-generated content available on the platform. Furthermore, the framework will serve as a foundation for future research and development in the field of sentiment analysis, facilitating advancements in sentiment classification techniques tai-

lored specifically for Twitter data.

1.1 Research questions

RQ1: To what extent does the inclusion of contextual information influence the results of LLMs in sentimental analysis tasks?

RQ2: To what extent does context play a role in making social media comments more trustworthy and reliable for humans?

1.2 Thesis outline

The following is the outline of the thesis:

- Chapter 2: Relevant background information, literature, and related work.
- Chapter 3: Methodology and research method used for this thesis
- Chapter 4: Collection of data, how it is used and results of the proposed conceptual framework. visual components, comparison of results to human survey. Before the research questions are discussed.
- Chapter 6. Conclusion and summary of what has been done. Recommendation for future work.

Objectives / Research questions

The primary goal of this master thesis is to explore ways of providing different perspectives on reactions to topics from individuals. The research will focus on analyzing certain posts related to specific events and determining the sentiment expressed in those posts in different ways. The conceptual framework will follow the common sentiment expressed in the posts and a modeling approach will be used to create a conceptual model that can detect sentiments better when contextual information is included. The interesting part of the research will be to compare the output of LLMs with human-generated social media data before and after the relevant context is added. Also, it will look for interesting tweets based on word clouds within a certain theme, which can be analyzed later using scripts.

1.3 Limitations

While this research provides valuable insights into the influence of contextual information on sentiment analysis and the preferences of participants, it is important to acknowledge its limitations. Some of the limitations of this research include the data set sample size. The research has relied on a relatively small sample size both of participants and data, which could limit the generalizability of the results. A larger and more diverse sample would provide a more comprehensive understanding of how people perceive and interpret sentiment in social media comments and detect emotions. Further, the participants involved in the survey may have exhibited inherent biases or subjective preferences that could influence their responses. Factors as personal experiences, cultural backgrounds, or individual perspectives may have affected their decision-making process and preference for contextual analysis. It should also be mentioned that the specific events or tweets selected for analysis might not represent a wide range of the chosen topics or contexts. The conclusions drawn from this research may be limited to the chosen events and may not be applicable to other types of events or different social media platforms.

Chapter 2

Background

Background literature and related work on sentimental analysis, large language models, and technologies relevant to this thesis are presented in this chapter.

2.1 Sentiment analysis

Sentiment analysis is the field of study that analyzes people's opinions, sentiments, attitudes, and emotions toward certain topics and their attributes expressed in written text [12]. Sentiment analysis techniques range from lexicon-based methods that rely on sentiment dictionaries to machine learning and deep learning models that leverage contextual information to classify sentiments. Researchers have developed sophisticated algorithms to analyze sentiments expressed in social media. It is a branch of natural language processing that focuses on identifying and extracting subjective information from text data. Unlike factual information, sentiment and opinion have important characteristics, namely, they are subjective. The subjectivity comes from many sources. First of all, different people may have different experiences and thus different opinions [12]. It aims to determine the sentiment or emotional tone expressed in a given piece of text, such as positive, negative, or neutral. By analyzing textual data, sentiment analysis offers valuable insights into public opinion, customer feedback, social media trends, and brand reputation.

The field of sentiment analysis has seen significant advancements and contributions from researchers, who have proposed various techniques and approaches to tackle this challenging task. This section will explore research areas and papers within sentiment analysis.

One common approach in sentiment analysis is lexicon-based methods, which rely on sentiment lexicons or dictionaries. These lexicons contain a collection of words or phrases labeled with their associated sentiment polarity. Researchers have developed and expanded

sentiment lexicons to improve the accuracy of sentiment analysis. There are three main existing approaches to compiling sentiment words: manual approach, dictionary-based approach, and corpus-based approach [12]. Where the manual approach is quite intensive and time-consuming, and is usually used as a check on automated approaches because automated approaches make mistakes.

Machine learning techniques have also played a crucial role in sentiment analysis. Researchers have explored various algorithms to automatically learn patterns and relationships between textual features and sentiment labels. Support Vector Machines (SVM) and Naive Bayes classifiers are commonly used in sentiment analysis tasks. Pang et al. [19] applied both Naive Bayes and SVM classifiers to movie reviews and achieved high accuracy in sentiment classification. They were using human-generated reviews because the domain is experimentally convenient due to large online collections of reviews, and because reviewers often summarize their overall sentiment with a machine-extractable rating indicator [19]. Where the author rating was expressed either with stars or some numerical value. By using features based on unigrams appearing at least four times in a 1400 document corpus, the work showed the effectiveness of machine learning algorithms in sentiment analysis. Compared with their random-baseline scoring of 50%, and the human-selected unigram baselines of 58% and 64%, their SVM score of 82.9% demonstrated both against the Naive Bayes classifier and compared to a 69% baseline achieved with limited access to test-data statistics competitive performance.

2.1.1 Natural Language Process

A field that contributes to continuous development within the field is Natural Language Processing (NLP). NLP is a field that combines linguistics, computer science, and artificial intelligence to enable computers to interpret, understand, and generate responses in human language. It covers a broad range of techniques, algorithms, and models that facilitate the processing and analysis of natural language text. NLP plays an important role in enabling machines to communicate with humans in a more intuitive and meaningful way.

NLP research focuses on various aspects of language processing, aiming to uncover the underlying structures, meanings, and patterns within text or speech data. One key area of research within NLP is syntactic parsing, which involves analyzing sentences. It means to break down a given sentence into its 'grammatical constituents' [26]. Syntactic parsing helps identify the relationships between words and phrases, providing insights into the syntactic hierarchy and syntactic roles of linguistic units.

Regarding the use of human-generated data from Twitter, Tang et al. [27] proposes a method

called sentiment-specific word embedding (SSWE) for sentiment analysis. They extend an existing word embedding learning algorithm and develop three neural networks to incorporate sentiment polarity supervision in their loss functions. They train their SSWE from a large collection of tweets that contain emoticons, serving as weakly supervised training data based on positive and negative emotions.

The SSWE is then used as a feature in a supervised learning framework for Twitter sentiment classification. Tang et al. [27] shows that the word embedding learned by traditional neural networks is not effective enough for Twitter sentiment classification. These methods typically only model the context information of words so that they cannot distinguish words with similar context but opposite sentiment polarity (e.g. good and bad). When evaluated on a benchmark data set, the method achieves competitive results with a macro-F1 score of 84.89% using only SSWE as a feature, comparable to the top-performing system using handcrafted features. By combining the SSWE feature with existing features, the state-of-the-art performance is further improved to 86.58% in macro-F1.

The quality of SSWE is also evaluated by measuring word similarity in the embedding space for sentiment lexicons, where it outperforms existing word embedding learning algorithms. The authors highlight three major contributions of their work: developing neural networks for learning SSWE from distant-supervised tweets, being the first to exploit word embedding for Twitter sentiment classification, and releasing the learned SSWE for adoption in other sentiment analysis tasks.

Another paper that is researching sentiments in Twitter data is Kiritchenko et al. [10]. They describe a supervised statistical sentiment analysis system designed for short informal textual messages like tweets and SMS. The system utilizes both existing general-purpose sentiment lexicons and tweet-specific lexicons generated from millions of tweets. These lexicons capture the peculiarities of social media language, including misspellings, elongations, and abbreviations.

Despite a lot of data and time-consuming tasks, the system processes tweets efficiently, where they highlighting the capability of analyzing 100 tweets per second. Where the ablation experiments demonstrate that the use of the automatically generated lexicons results in performance gains of up to 6.5 absolute percentage points. [10].

2.1.2 Word to vector

Converting words to vectors, or word vectorization, is a NLP process. It can be described as using "language models to map words into vector space. A vector space represents each word by a vector of real numbers. It also allows words with similar meanings to have simi-

lar representations [11]. By capturing their semantic relationship, it employs shallow neural networks to learn distributed word representations based on the context in which words appear. This approach enables models to capture similarities and associations between words, facilitating tasks such as word analogy, word similarity, and language understanding.

The advancements in machine learning techniques, particularly the use of distributed representations of words, and their impact on improving language models are discussed in Mikolov et al. [14]. They aim to introduce techniques for learning high-quality word vectors from large data sets with billions of words. By evaluating the quality of the resulting vector representations using measures that capture both similarity and multiple degrees of similarity between words.

Some of their highlights are word representations exhibit similarities beyond simple syntactic regularities [14]. They demonstrate this by performing algebraic operations on word vectors, such as using a word offset technique where simple algebraic operations are performed on the word vectors, it was shown for example that vector(”King”) - vector(”Man”) + vector(”Woman”) results in a vector that is closest to the vector representation of the word Queen [14].

Conducting experiments to compare the quality of vector representations from different models using a range of syntactic and semantic language tasks, they find that simple model architectures can produce high-quality word vectors, outperforming more complex neural network models. The computational complexity of these simpler models allows for the computation of accurate high-dimensional word vectors from significantly larger data sets.

They conclude that earlier work on unsupervised vector-based approaches to learning word vectors for sentiment analysis [13], can expect that these applications can benefit from the model architectures described in [14]. Additionally, ongoing work suggests the successful application of word vectors to tasks like extending facts in knowledge bases, verifying the correctness of existing facts, and machine translation experiments.

2.2 Large Language Models

Large Language Models (LLMs) represented a milestone in NLP research and development. LLMs are powerful models that have been trained on massive amounts of textual data, enabling them to understand and generate human-like text. These models leverage advanced techniques such as deep learning, and attention mechanisms.

Devlin et al. [3] introduced a language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. BERT is designed to pre-train

deep bidirectional representations by considering both the left and right context in all layers. This allows the pre-trained BERT model to be fine-tuned with an additional output layer, enabling it to achieve state-of-the-art performance on various tasks without extensive modifications to the task-specific architecture.

BERT is both conceptually simple and empirically powerful. It surpasses previous models and achieves new state-of-the-art results on eleven NLP tasks. It achieves an 80.5% GLUE score (General Language Understanding Evaluation benchmark) which is a 7.7% absolute improvement. An 86.7% MultiNLI accuracy (a 4.6% absolute improvement), a 93.2 SQuAD (The Stanford Question Answering Dataset) v1.1 question answering Test F1 (a 1.5 point absolute improvement), and an 83.1 SQuAD v2.0 Test F1 (a 5.1 point absolute improvement). BERT demonstrates its effectiveness in a range of language-related tasks, showcasing its capabilities and advancements in the field [3].

By employing a deep bidirectional transformer architecture and pre-training on large corpora to capture contextual information effectively. By leveraging masked language modeling and next-sentence prediction objectives, BERT learns rich representations that excel in a range of NLP tasks, such as sentiment analysis.

LLMs are pre-trained on large-scale corpora, utilizing unsupervised learning to learn the statistical properties and contextual relationships of words and phrases. By training on vast amounts of text data, LLMs develop a rich understanding of grammar, semantics, and pragmatic aspects of language. This pre-training phase is followed by fine-tuning specific downstream tasks, where the models are adapted to specific domains or applications.

Research in the field of LLMs continues to advance rapidly. In the next section 2.2.1, we will discuss state-of-the-art models, how these LLMs are developed, and background information on why it will be used for further investigation in this thesis.

2.2.1 Transformer

The Transformer architecture, first introduced in the paper Vaswani et al. [28], utilizes self-attention mechanisms to capture contextual relationships between words or tokens in a sequence efficiently and "is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution."

Traditional NLP models heavily relied on recurrent neural networks (RNNs) like LSTM and GRU, which process sequential data sequentially, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples Vaswani et al. [28].

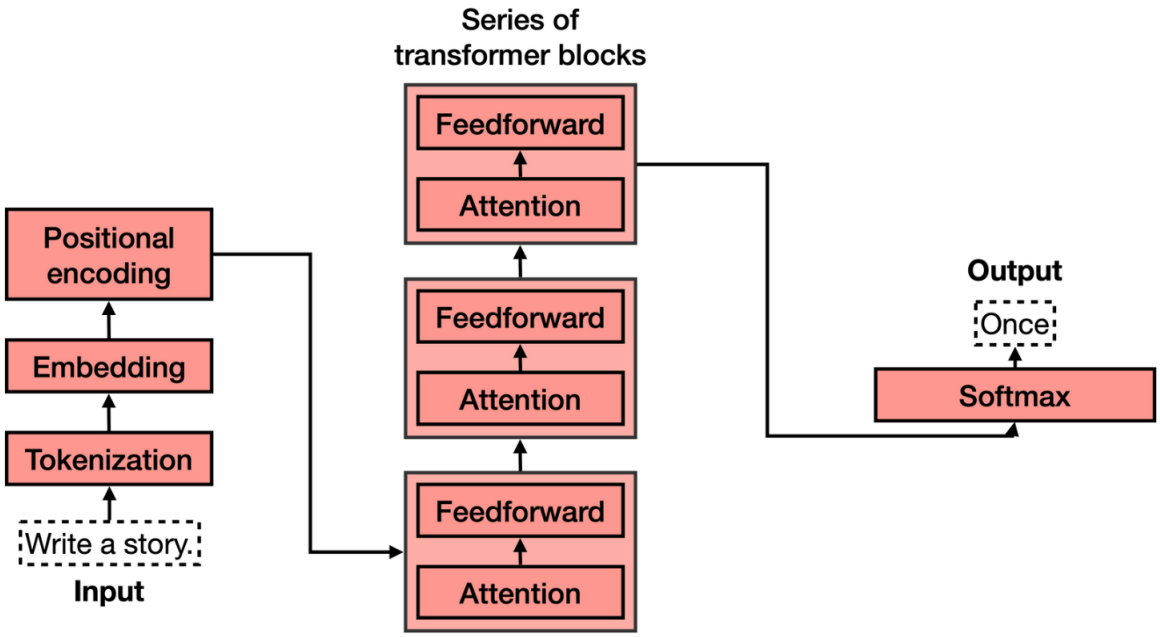

Figure 2.1: The architecture of a transformer model [24]

A transformer model consists of several parts. The fundamental concept in Transformers is the attention mechanism, shown in 2.1, which computes attention weights to determine the importance of each token with respect to other tokens in the sequence. Combined with a feed-forward neural network within each layer, called a series of Transformer blocks, it applies non-linear transformations to the representations of tokens, enabling the model to capture complex patterns and relationships within the sequence. The specific design of the feed-forward network, including the number of layers and hidden units, can impact the model's performance and computational efficiency.

Before the attention and feed-forward layers, 2.1 shows tokenization, embedding, and positional encoding. Tokenization is an initial step in NLP, where every word, prefix, suffix, and punctuation sign, and sends to a known token from the library [24]. Once the text is tokenized, an embedding process is applied to convert words into numerical representations. Embeddings play a crucial role in language models as they serve as a bridge between text and numbers. They transform the text into vectors of numbers, where similar texts have similar vector representations and dissimilar texts have different vector representations [24]. This geometric interpretation allows us to visualize embeddings as points in a coordinate system, called positional encoding. Positional encoding first proposed in [25], presented an extension to self-attention that can be used to incorporate relative position information for sequences, which improves performance for machine translation. Similar words cluster together and dissimilar words are located farther apart. For instance, in a simplified embedding with 2-dimensional vectors, the word "cherry" would have coordinates close to "straw-

berry" but far from "castle" [24].

The final step in 2.1 is the softmax layer. After predicting the next word in a sentence, the softmax layer turns these scores into probabilities, where the highest scores correspond to the highest probabilities [24]. This probability distribution allows us to sample the next word based on the predicted probabilities.

Transformer architecture has advanced NLP research and achieved state-of-the-art performance in various benchmarks. One of the most influential applications and widespread knowledge of Transformer architecture is the development of language models. OpenAI's Generative Pre-trained Transformer (GPT), has demonstrated exceptional performance in various NLP tasks. The GPT model was introduced by Radford and Narasimhan [20], which showed the effectiveness of pre-training and fine-tuning strategies with Transformers. By training a Transformer-based language model on a large corpus of diverse text data, GPT leveraged unsupervised learning to learn powerful language representations.

GPT follows two-step training process commonly used in Transformer models. During pre-training, the model learns to predict missing words in sentences, similar to masked language modeling, as well as to predict the next word given the previous context. This pre-training phase allows GPT to capture extensive knowledge about language structure and semantics. After pre-training, GPT is fine-tuned on specific downstream tasks using supervised learning. By fine-tuning the pre-trained model on labeled data from specific tasks, GPT adapts its learned representations to the target task. The GPT-3 in the series of development described in the paper Schick and Schütze [23] shows the power of large-scale Transformer models. With 175 billion parameters, GPT-3 achieved unprecedented performance on a wide range of NLP tasks, including language translation, question answering, and even creative writing. The success of GPT-3 demonstrates the significant potential of LLMs based on Transformer architectures. These models have become the focal point of NLP research, attracting attention from both industry, academia and people without special knowledge of machine learning. Their ability to generate human-like text and perform complex language tasks with minimal task-specific training has opened up new possibilities in various domains.

Therefore, the GPT-3 will serve a crucial role in the development of the conceptual framework in this thesis.

2.2.2 Fine tuning LLMs

Fine-tuning is a process that allows us to influence the power of Transformer models for specific NLP tasks. It involves adapting a pre-trained model, which has learned from a vast amount of unlabeled text data, to a specific task using a smaller labeled data set. By fine-

tuning, we enable it to transfer learned knowledge to the target task, resulting in improved performance and task-specific capabilities.

Pre-training: Initially, the Transformer model is pre-trained on a vast amount of unlabeled text data using unsupervised learning. This pre-training phase involves tasks such as masked language modeling, where the model predicts missing words in sentences, and next sentence prediction, where the model predicts whether two sentences follow each other coherently. This pre-training process enables the model to learn contextual representations and capture various language patterns and semantics [3].

Task-specific Data set: After pre-training, the model is fine-tuned on a smaller, task-specific dataset that is labeled. This dataset contains examples relevant to the target task, such as sentiment analysis or named entity recognition. The labeled data allows the model to learn task-specific patterns and refine its representations.

Architecture and Parameters: During fine-tuning, the architecture of the pre-trained Transformer model remains intact. However, some task-specific layers or components may be added on top of the pre-trained model to adapt it to the specific task requirements. The parameters of the pre-trained model, including the attention weights, positional encodings, and feed-forward networks, are updated based on the labeled task-specific data.

The fine-tuning process has proven effective due to the transfer learning capabilities of Transformer models. Pre-training on large-scale unlabeled data enables the models to learn general language representations while fine-tuning on task-specific data tailors these representations to the specific task requirements. This approach reduces the need for large labeled datasets, as the models can leverage the knowledge learned during pre-training.

Fine-tuning Transformer models has become a standard practice in many NLP applications, allowing researchers and practitioners to harness the power of pre-trained models for various downstream tasks. The process enables efficient transfer of learned representations and significantly improves performance, making Transformer models highly versatile and applicable across a wide range of NLP domains.

2.2.3 Hallucination problem

Hallucination is a significant problem within sentiment analysis, particularly when utilizing LLMs without proper awareness. With the mentioned GPT-3 model with ability to understand and generate human language, it does not always retrieve the correct information. Hallucination refers to the generation of false or fabricated information by the model, in our case leading to inaccurate sentiment analysis results. It occurs when the LLMs generates text that appears coherent but is not based on actual facts or sentiments expressed in the input.

When people use LLMs for sentiment analysis without recognizing the potential for hallucination, they may unknowingly rely on the generated outputs, assuming them to be accurate and trustworthy.

For both the data collection and results the following problems could occur in different ways:

Misleading insights: Hallucination can distort the interpretation of sentiment analysis results, leading to incorrect conclusions and decision-making. If the LLMs fabricates positive or negative sentiments that do not align with the actual text, it can provide misleading insights and misrepresent the sentiment expressed in the data.

Biased outputs: LLMs are trained on large corpora, which can contain biases present in the data. If the LLMs hallucinates sentiments based on biased patterns in the training data, it can perpetuate or amplify existing biases in sentiment analysis. This can lead to unfair judgments and reinforce stereotypes or discrimination.

Trust issues: Users who are unaware of hallucination in LLMs may develop a false sense of trust in the generated outputs. They may rely heavily on the model's results without critically evaluating or verifying them, potentially leading to flawed analyses and misguided decisions.

A challenge associated with LLMs and their unintended behaviors, such as generating biased or toxic text, not following user instructions, or making up facts is discussed in Ouyang et al. [17]. The language modeling objective used for many LLMs, which involves predicting the next token on a web page from the internet, is misaligned with the objective of following user instructions helpfully and safely. Aligning language models with user intentions is crucial, especially considering their widespread deployment in various applications.

By proposing a fine-tuning approach called reinforcement learning from human feedback (RLHF) to align language models, focusing on GPT-3. They employ human preferences as a reward signal to fine-tune the models through a multi-step process. Initially, a team of contractors labels the data based on their performance on a screening test. Human-written demonstrations of desired output behavior are collected, along with comparisons between model outputs on a larger set of prompts. A reward model (RM) is trained on this data set to predict preferred model outputs, and the supervised learning baseline is fine-tuned using the Proximal Policy Optimization (PPO) algorithm.

The resulting models, called InstructGPT, align the behavior of GPT-3 with the preferences of the labelers and researchers involved in the process. The evaluation of the models includes rating the quality of model outputs by the labelers on a test set consisting of prompts from held-out customers. Additionally, automatic evaluations are conducted using various public NLP data sets.

The approach outlined in the text addresses the misalignment of language models' objec-

tives and aims to make them helpful, honest, and harmless. By fine-tuning the models through RLHF, they strive to improve the models' behavior in accordance with user intentions and mitigate unintended behaviors. The evaluation process involves both human assessments and automatic evaluations on diverse data sets, where despite the strong performance of the 175B PPO-ptx language model in various language tasks, it still exhibits certain limitations and can make simple mistakes in certain situations [17].

2.2.4 Detecting emotions from text

The detection of emotions within sentiment analysis poses several challenges due to the complexity and subjectivity of human-generated content. While sentiment analysis focuses on classifying text as positive, negative, or neutral, emotions delve into more nuanced and multi-dimensional aspects of human affective states. An Ontology-Based Sentiment Analysis approach, which uses a taxonomy of emotional concepts to extract emotions from tweets related to COVID-19 was done by Narayanasamy et al. [15]. The approach analyzes potential entities in the tweets for semantic associativity. To extract the emotions, they applied the a SVM classifier to effectively identify eight emotions. In 2.2 the Domain Ontology was created manually with the help of a psychological-emotional words list and WordNet. This gives a picture of both the amount of emotions and how difficult it could be to identify and choose in which category a human-generated text fits into. In that way, they demonstrated a wide range of human emotions specifically related to COVID-19 concepts. By adding subsets of emotions to the eight primary emotions, they classified emotions into four types. Basic Emotions (joy, confusion, anger, disgust, fear, sadness, love, and surprise), Mid Emotions (distraction, boredom, acceptance, apprehension, interest, serenity, pensiveness, and annoyance), Intense Emotions (ecstasy, amazement, vigilance, grief, admiration, loathing, rage, and terror) and Complex Emotions (disproval, love, submission, optimism, awe, contempt, aggressiveness, and remorse) [15].

The construction of the ontology involved the use of Semantic Web technologies, with the creation of triples for each tweet and manual implementation of Web Ontology Language (OWL) functions. They found that the extraction of emotions can be performed using words directly or indirectly associated with emotional concepts in the taxonomy and was was able to detect 81% of true positives and is able to detect a considerable amount of false negatives, within the English language. The reason for showing an ontology with a focus on emotion is to give a picture of how hard it can be not to just group sentiments in the general groups, positive, negative, and neutral. It is a difficult task to detect sentiment, but as shown, emotions that occur in different groups and for different purposes can describe the same feelings. This is relevant for later analysis when a comparison of contextual and non-contextual sentiment

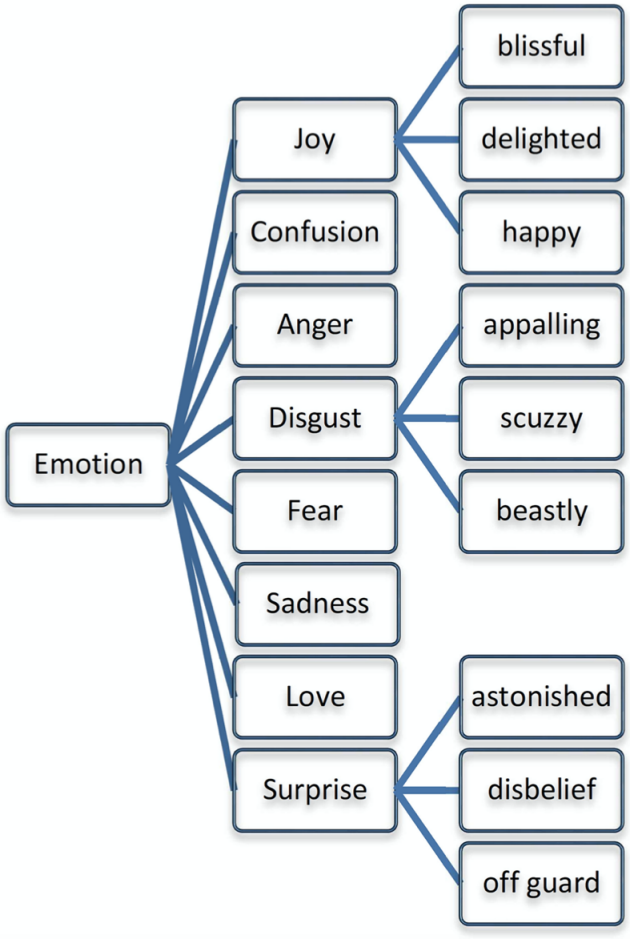

Figure 2.2: Fundamental emotion list used for ontology population [15]

analysis is conducted.

2.3 Sentiment Analysis for Twitter

The rise of social media platforms has led to increased interest in sentiment analysis for social media data. Researchers have focused on developing techniques to handle the unique characteristics of social media language, including abbreviations, slang, and informal expressions. According to the report of USD dollar in the social media market [4], the increasing demand for social media monitoring and the growing use of sentiment analysis in customer experience management and branding as some of the key factors driving the market growth.is conducted on sentiment analysis of Twitter data and compared lexicon-based and machine-learning approaches. Their work demonstrated the challenges and opportunities in analyzing sentiment in a social media text.

The goal of sentiment analysis in the Twitter domain is traditionally to classify text into three categories: positive, negative, or neutral. To do this, machine learning algorithms are used to analyze various features of the text, such as the words used, the sentence structure, and the context in which the text appears. Pak and Paroubek [18] focuses on sentiment analysis and opinion mining in microblogging platforms, specifically Twitter. The paper shows how to automatically collect a corpus for sentiment analysis purposes, perform linguistic analysis of the corpus, and build a sentiment classifier that is able to determine positive, negative, and neutral sentiments for a document. The paper also discusses the discovered linguistic phenomena and evaluates the proposed techniques, which are shown to perform better than previously proposed methods.

Therefore, as a combination of factors, with the rise of availability and accessibility of AI for the population in total, the awareness and importance of trustworthiness when relying on AI needs to be further investigated and discussed. With this provided background information, and in the field of analyzing sentiments from Twitter, we can not see that people use LLMs and detect emotions from human-generated data on Twitter. The following sections will suggest a conceptual framework and show how context can be provided to extend the sentiment analysis in the Twitter domain.

Chapter 3

Methodology

This chapter is describing the methodology and methods used in the research.

3.1 Design science research

Design science is a problem-solving process of designing artifacts to solve problems. Design science focuses on the relevance of IT artifacts in applications. [6] defines DSR as follows: "Design science research is a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence. The designed artifacts are both useful and fundamental in understanding that problem."

In this section, we will explore the DSR in more detail, discussing its steps, advantages, limitations

To provide an understanding for researchers and other stakeholders regarding the requirements of effective design science research, a guideline of in total seven where created by [7]. The first guideline is *Design as an artifact* and states “Design- science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiating.” [7]. By this, it means that the research you are doing should contribute, be purposeful, and be described effectively. The thesis’s main purpose will be to visualize and show if contextual information can provide a better understanding of finding interesting tweets based on tweets that stands out from the rest and perhaps shows a different side of the theme or event. The research aimed to design and develop a framework for analyzing sentiment in social media comments with and without context. The framework was treated as an artifact, consisting of various components and transformations to incorporate contextual information into sentiment analysis models.

Table 1. Design Science Research Guidelines	
Guideline	**Description**
Guideline 1: Design as an Artifact	Design-science research must produce a viable artifact in the form of construct, a model, a method, or an instantiating,
Guideline 2: Problem Relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Guideline 3: Design Evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Guideline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research Rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a Search Process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audience.

Table 3.1: The guidelines in design science research from Hevner et al. [7]

Identifying, defining the problem, and understanding its context is necessary. It is important to have a clear understanding of the problem to ensure that the solution developed addresses the root cause of the problem. This step also involves understanding the stakeholders involved in the problem, such as users and customers. The research identified a relevant problem in sentiment analysis, specifically focusing on the influence of contextual information on the accuracy and reliability of sentiment analysis results. This problem is of practical importance, as sentiment analysis plays a crucial role in understanding public opinions and attitudes on social media platforms.

The next guideline is *Problem relevance* which is defined: "The objective of design-science research is to develop technology-based solutions to important and relevant business problems." [7]. Where you are at your current state and the difference from what you define as your goal, is what´s called the problem you are trying to solve. For design science, relevant business problems are addressed by the construction of innovative artifacts [7].

The third guideline is *Design evaluation*. It is defined as "The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods" [7]. The types of evaluation methods in design science are observational, analytical, experimental, testing, and descriptive. To be able to evaluate the research in the right way are very important if the artifact should be able to contribute to the field. The evaluation of the designed artifact was carried out through a combination of qualitative and quantitative methods. The research used surveys and participant feedback to gather insights on the perceived accuracy,

trustworthiness, and reliability of sentiment analysis results with and without context. The evaluation also considered the preferences and perspectives of human readers when interpreting social media comments.

The fourth guideline is *Research contributions*. It is defined as Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies" [7]. The artifact may extend the knowledge base or apply existing knowledge in new and innovative ways. The research contributes to the existing body of knowledge by investigating the role of context in sentiment analysis. It provides insights into the impact of context on the accuracy and reliability of sentiment analysis outcomes and highlights the significance of incorporating contextual information in sentiment analysis models.

The fifth guideline is *Research rigor*. It is defined as "Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact." Rigor is achieved by effective use of the foundations and research methodologies from the knowledge base. Research methods from the literature will be followed throughout the research process. The research followed rigorous scientific methods by employing a systematic approach to data collection, analysis, and interpretation. It used a controlled experimental setup where participants were presented with social media comments both with and without context and their responses were recorded and analyzed. Statistical analysis was performed to identify patterns and trends in the data.

The sixth guideline is *Design as a search process*. It is defined as "The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment." [7]. The research involved iterative design and refinement processes. The initial framework was developed based on existing literature and theories in sentiment analysis. Feedback from participants and domain experts was collected and used to refine the framework, ensuring that it aligns with the research objectives and addresses the identified problem effectively.

The seventh guideline is *Communication of the research* and states "Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences." [7]. The findings of the research were communicated through various means, including academic papers, conference presentations, and this chat. The research outcomes, methodology, and implications were clearly articulated to facilitate knowledge sharing and foster further discussions in the field.

By adhering to these seven guidelines for design science research, this study ensures a rigorous and systematic approach to investigating the role of contextual information in sentiment analysis. The methodology employed enables the research to make meaningful contribu-

tions to the field and provide valuable insights for future research and practice.

3.1.1 Use of generativ AI

A rapidly evolving field that has garnered significant research attention in recent years is Generative AI. It has a profound impact on various fields, and one notable application is in NLP. LLMs like OpenAI's GPT have demonstrated remarkable text generation capabilities, while the primary purpose of LLMs is to generate coherent and contextually relevant text, researchers and practitioners have also explored their potential in sentiment analysis, including the analysis of sentiments expressed in tweets.

GPT, with its impressive language generation abilities, can be harnessed to detect sentiment from tweets through the use of prompts. Prompts are specific instructions or cues provided to the model to guide its output. By framing a prompt that encourages the generation of text expressing sentiment, it is possible to leverage GPT for sentiment analysis tasks.

In the context of sentiment analysis on tweets, one approach is to provide a prompt that explicitly instructs the model to generate sentiment-related content. For instance, a prompt such as "Please generate a tweet expressing positive sentiment about [topic]" can guide GPT to generate text that reflects a positive sentiment regarding the given topic. By analyzing the generated text, sentiment can be inferred based on the expressed emotions and opinions.

It is important to note that GPT is not specifically designed for sentiment analysis, and the generated output should be treated as a starting point for further analysis and validation. Post-processing steps, such as filtering out irrelevant or nonsensical content and applying sentiment analysis techniques, may be necessary to obtain accurate sentiment classifications. In summary, generative AI models offer opportunities for sentiment analysis by utilizing prompts to guide the generation of sentiment-related text.

3.1.2 Enrichment with contextual information

With a conceptual framework for sentiment analysis using LLMs on social media data. Employed with a design science research methodology, testing and refining the proposed framework through iterations, the focus is on detecting the importance of context in improving the reliability of sentiment analysis. Each step in the framework has a descriptive text and hopefully, results in the box "Common sentiments to topic". Here each topic, "Topic 1", "Topic 2" and are, as we will see later, have the prejudge in which direction the feelings will be. When this is established and a new emotion is collected which has not occurred before. Becuase this is a "zero-shot", the ability to perform tasks without any specific training examples or ex-

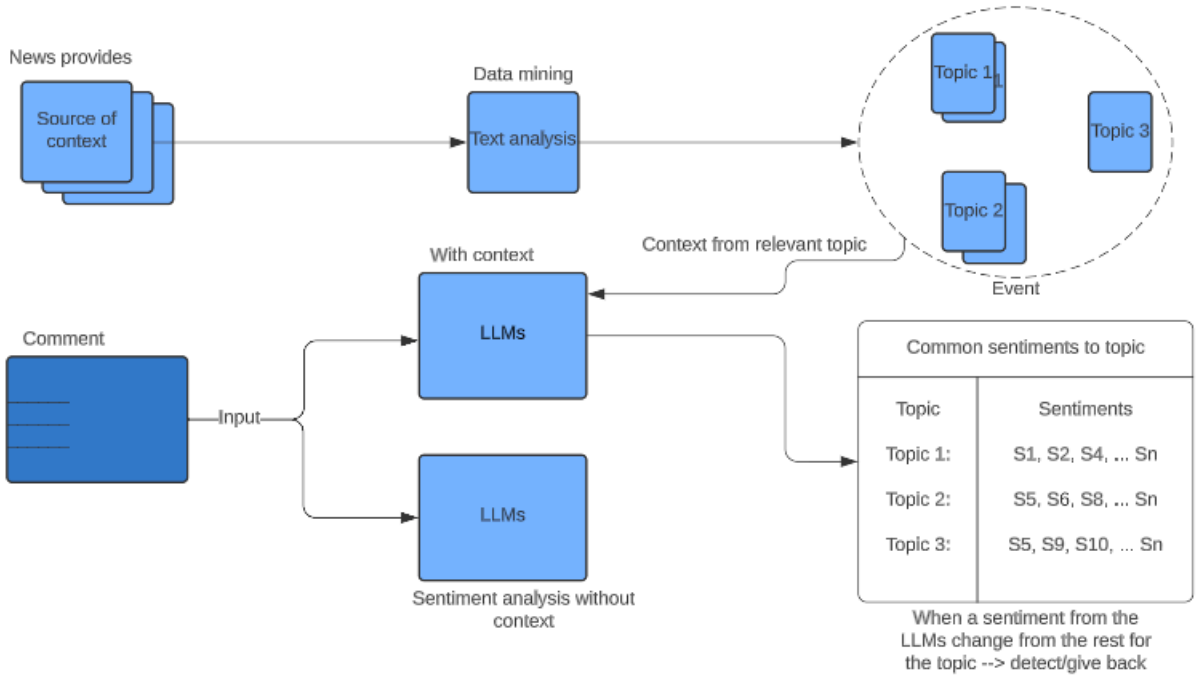

Figure 3.1: Proposed framework of detecting emotion from Tweets

plicit supervision for that particular task is provided. It means that the model can generalize to new, unseen tasks based on its understanding of the underlying language and the knowledge it has acquired during training. With zero-shot learning can be useful when dealing with tasks or domains for which training data is limited or unavailable. Instead of relying on task-specific training data, a zero-shot model leverages its broader language understanding and general knowledge to perform the task.

The proposed framework is evaluated using a human survey, comparing sentiment analysis with and without the incorporation of contextual factors, as shown in 3.1. The framework has the potential to improve the use of sentiment analysis in various domains, such as political analysis and journalistic work and for people using the more and more available technology.

Chapter 4

Results

This chapter will present the data, the collection, and how it is used. Introduce topic mining and how other technologies can be used on contextual data. Then a review of the survey and how it contributes, before analyzing and evaluating the results regarding the research questions.

We have addressed the problems regarding detecting sentiments in text. Anomaly detection in sentiment analysis is an aspect that involves identifying abnormal opinions, sentiment patterns, or temporal aspects within a data set. These anomalies can be attributed to sudden shifts in sentiment concealed within extensive text data. Failing to detect or effectively manage these anomalies can lead to severe consequences, such as negative sentiments expressed by customers. Twitter serves as a vast information source information on various topics. Hence, analysis of social media data to identify abnormal events holds significant value, for example by allowing to intervene early or adopt appropriate strategies when necessary for businesses. However, challenges due to the diverse nature and large volume of social media data can be hard. In Wang et al. [29] anomaly detection in sentiment analysis is discussed. The study explores existing anomaly analysis and sentiment analysis methods, highlighting their limitations and challenges. To address these challenges, an enhanced sentiment classification method is proposed and discussed. The researchers examine the potential of using this method for anomaly detection through sentiment analysis on social media data, specifically tweets. The results demonstrate the effectiveness and robustness of the proposed method, offering valuable insights into this research domain.

For the purpose of collecting, analysing and present sentiments and emotions regarding different topics and events. In the next section, the work behind the presented framework will be represented and explained in more detail.

Regarding RQ2, to what extent does context play a role in making social media comments

more trustworthy and reliable for humans? 1.1, it is important to show a wide range of the population. By presenting a range of opinions and perspectives, for example, humans can gain a deeper understanding of complex issues and events. For example, this could help journalists and news providers help the audience gain a deeper understanding of complex issues and events. They can also encourage critical thinking and informed debate, allowing readers and viewers to form their own opinions based on a variety of viewpoints.

4.1 Prompt engineering

Prompt engineering refers to the process of designing or selecting input formats, consisting of natural language descriptions and task-specific instructions, to enable a language model to perform a specific task. With prompt engineering, it aims to improve the quality of the model's output by enhancing the relevance of the input text to the task at hand.

The technique is used to improve the performance of NLP models and provide a more preferable outcome. In general, it involves designing or selecting the natural language input that enables the model to perform a specific task. Because the models often are trained and tested on an enormous amount of data, one of the key advantages of prompt engineering is that it enables models to generalize better to new tasks and domains. The prompts can be tailored to the specific requirements of the task. Additionally, prompt engineering can reduce the amount of labeled data required for training. If the prompts are designed with the necessary information, a benefit could be that the model can be trained on a smaller corpus.

The contextual approach will use the OpenAI´s GPT-3 [16], described in 2.2.1. Although this model is not specifically designed for sentimental analysis it will be used to process words and responses via prompts. The combination of the research questions provided 1.1, and since we're investigating words that are associated with human emotions, and not numbers, the wide range of textual data it is trained on, and the ability to generate coherent responses on given prompts, will give good insights when the research questions are taken into consideration. Because the model allows users to create prompts that guide the model to generate text that follows a specific pattern or theme. GPT-3 will serve on sentiment analysis, without fine-tuning by providing it with labeled data and training it to generate relevant output based on specific prompts. This is because the main research will conduct by comparing how context can affect the output.

There are many different prompts you can use with GPT-3, depending on your goals and the type of text you want to generate. In general, examples of prompts can be:

- Generating a description of a product: "Write a paragraph describing the features and

Generate a sentence describing the feelings that the [text] contains:

Table 4.1: The prompt used for sentiment analysis on Twitter data

Prompt	Generate both a sentence and keywords of feelings describing the feelings that the comment contains over the context of the news given below:
Context	News: [Context for the specific event]
Data	Comment: [Tweet]

Table 4.2: The prompt used for contextual sentiment analysis on Twitter data

benefits of the product, [Product Name]."

- **To generate creative writing:** "Provide five prompts for a short story about a car stuck in traffic."

For the non-contextual approach, the following prompt was given: This prompt was universal and where applied to every tweet in the data set. The text field in 4.1 is the variable for input of each tweet.

When designing a prompt to take the contextual information available into consideration, we adopt a prompt engineering approach that incorporates the context before it was presented the comment from Twitter. The prompt structure for emotion classification with context was used as followed: The objective was that the model generates both a sentence and the associated keywords that described the sentiment expressed in each comment. To achieve this, there were intentionally provided minimal additional information to avoid confusion. However, it is important to note that the sentiment results may vary each time the prompt is given to the model due to the inherent variability of language models. Because of this, with the implementation of sentiment detection, the general text template would be common to all tweets. This template served as the basis for generating the sentence and keywords associated with each comment. By maintaining a consistent structure, we aimed to standardize the sentiment detection process across different tweets and topics and minimize the biases introduced by varying text contexts and other knowledge provided at the moment of analysis.

4.2 Contextual information

Context can play a vital role in understanding the sentiment expressed in text, for humans it provides the necessary background and situational cues for proper interpretation and understanding. As the research aims to investigate how trustworthy and impacted the output can be, the importance of providing correct, reliable, and accurate context is crucial. Even more important is to provide correct, and good contextual information when the purpose is to retrieve emotions via sentiment analysis. Trying to balance a fine line between incorpo-

Tweet	Description without context	Description with context
"Big miracle after 128 hours, a 2 month old baby was rescued from the earthquake rubble in Turkey. Hoping it makes a quick recovery."	The text contains feelings of hope and happiness.	The commenter expresses hope and relief after hearing about the miraculous rescue of a 2-month-old baby from the earthquake rubble in Turkey, and wishes for the baby's quick recovery.

Table 4.3: Result after both approaches of sentiment analysis to a tweet

rating context information and avoiding the use of explicit emotional words or cues. Striking the right balance is crucial to ensure accurate and in this matter unbiased sentiment analysis results.

Contextual information can be derived from various sources and in this research from the surrounding text. We try to avoid any previous dialogue, metadata associated with the text, and time, location, or subject matter. Incorporating this information helps capture the nuances and subtleties of the desired basis for comparison, and in that way leads to more accurate analysis.

It is important to avoid relying on emotional words or explicit cues to determine sentiment. Emotional words are subjective and can vary across individuals or cultures. Relying on emotional words may introduce biases or limitations in sentiment analysis, leading to inaccurate results. By explicit focus on factual information, the analysis may be able to better use the information it is provided to take its analysis.

With two main approaches: contextual-based and non-contextual sentiment analysis, both approaches have their strengths and weaknesses, in this section, the primary focus is on the contextual approach and how it is assigned.

As mentioned in 2.2.3, it should be difficult to fully trust the results or answers provided by the prompt and model.

4.2.1 Context from Reuters

Reuters is often considered a reliable source for obtaining context information due to several factors that contribute to its credibility and reputation. As one of the world's largest and most renowned news agencies, Reuters has established itself as a trusted source of factual and unbiased news coverage. Here are some reasons why Reuters is often preferred as a reliable source for context information because of the following:

Journalistic Standards: Reuters adheres to high journalistic standards and principles, including accuracy, integrity, and impartiality. The organization follows a strict editorial policy that emphasizes fact-checking, verification, and multiple sources to ensure the accuracy and reliability of its news reporting.

Global Coverage: Reuters has a vast global network of journalists and correspondents who

provide coverage from various regions around the world. This extensive coverage allows Reuters to capture diverse perspectives and provide comprehensive context on a wide range of topics and events.

Objectivity and Impartiality: Reuters strives to present news in an unbiased and impartial manner, focusing on factual reporting rather than promoting any specific agenda or bias. The organization maintains a commitment to neutrality, ensuring that the information it provides is free from personal opinions or editorial bias.

Reputation for Accuracy: Reuters has a strong track record of delivering accurate and reliable news content. It invests significant resources in fact-checking and verifying information before publishing it. Their commitment to accuracy has earned them the trust of readers, researchers, and professionals across various industries.

Established Editorial Guidelines: Reuters has well-defined guidelines that journalists follow when reporting the news. These guidelines emphasize the importance of balance, fairness, and accuracy that is provided in a framework for journalists to ensure that the context and information presented in their reports are reliable and trustworthy.

Transparency and Accountability: Reuters is committed to transparency and accountability in its reporting. In cases where errors or mistakes occur, the organization promptly corrects them and issues public clarifications. This commitment to transparency enhances its credibility and reliability as a source of context information.

As mentioned earlier, one of the key parts is to provide the model with mostly factual information. To do so, it was necessary to collect this information after selecting Reuters as the main source of information about events. Reuters' posted its most important happenings in February 2023 [22]. The content and themes covered by Reuters during that period determine the specific topics and keywords for what was impacting the world at that period of time. This forms the basis for 4.1, which includes a wide range of events and news stories such as global affairs, politics, economics, technology, science, sports, and more. The word cloud highlights significant events and keywords that dominated the news cycle during that particular time. The data is manually downloaded from the website and is further described under 4.3. It can be possible to select which type of events, period, and more for what type of event you would look deeper at sentiments being expressed by Twitter users.

4.3 Topic mining

The topic mining is planned and related to be used in different parts of the process. The . Therefore, there will be given an introduction to how it is used, and later of how as described

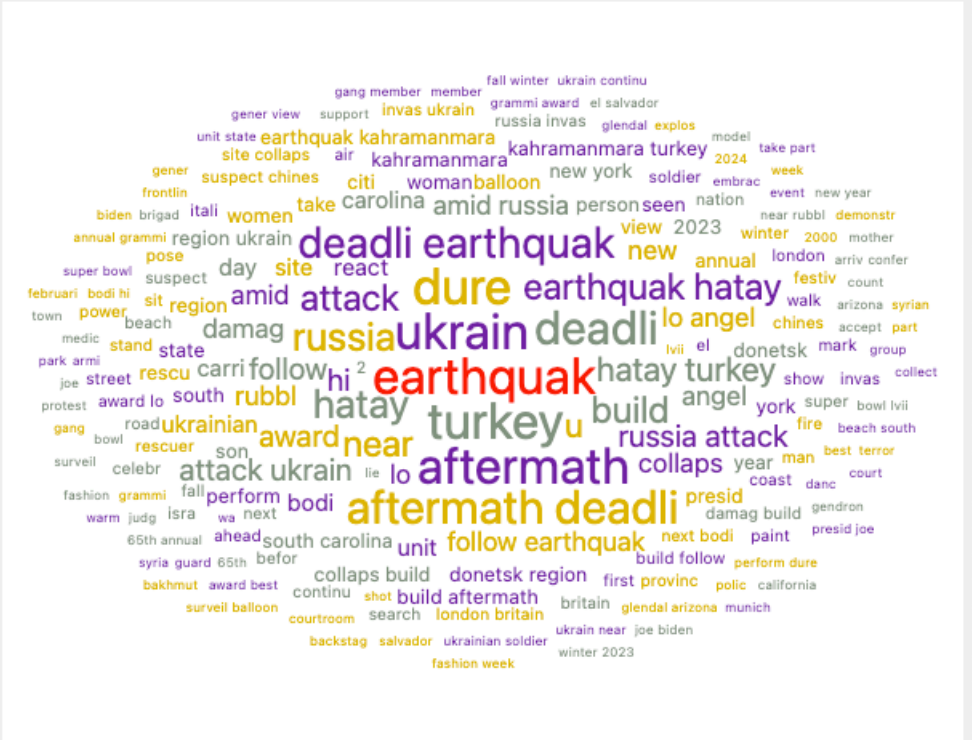

Figure 4.1: Summary of important events from Reuters in Feb 2023

in the figures

In 4.1, Orange3 is used as a topic mining tool to represent words that are often expressed in the data provided. Orange3 is an open-source data mining and visualization tool that offers a wide range of functionalities for exploratory data analysis, machine learning, and data visualization. One of its features is topic mining, which allows users to extract meaningful topics or themes from textual data.

It aims to uncover the underlying structure and patterns within the text, enabling researchers to gain insights into the main subjects or ideas discussed in the documents. For this research, the underlying meaning of why and how it performs as it does is not given any particular notice. It is used to see the potential for further development and use, and at the same time give the proposed framework some visual parts.

Word clouds are a popular and visually appealing method for representing textual data. They provide a concise and intuitive way to visualize the frequency or importance of words within a given text corpus. Here are some of the benefits of using word clouds:

It gives a quick overview of key terms. An efficient overview of the most frequent or prominent terms in a text corpus. The size of each word in the cloud corresponds to its frequency or relevance, in this case, both the possibility to determine important news 4.1 or the sentiment that occurs within an event. This can be particularly useful for gaining a high-level understanding of the main topics or themes within a large body of text.

Further, it can identify patterns and trends within the text by highlighting recurring terms. Identifying Patterns and Trends: Word clouds can reveal patterns and trends within the text by visually clustering related words together, word clouds allow users to identify common topics or concepts that emerge from the text corpus. This can aid in uncovering hidden insights, understanding prevalent sentiments, or identifying key terms that are frequently associated with a specific domain or context. And to use the coluds as an ecploratory starting point for further analysis. For example, when identifying words or that occur, but not that many times can be selected to look further into.

Orange3 provides a user-friendly interface and a set of pre-built widgets that facilitate topic-mining tasks. These widgets allow users to preprocess textual data. The first step in **??**, is to choose the desired corpus. Then you can select which rows/columns you would like to focus on. With this step, you are able to change the visualization from the context based data to the non-contextual data in a easy way. "Preprocess Text" involves transforming the textual data into numerical representations that capture the semantic meaning and relationships between documents. This step facilitates clustering and similarity analysis and is based on default settings in Orange3. Enabling the opportunity for groups or clusters of related keywords based on their content.

The data can now be visualized. By using a distance metric on a hierarchical clustering technique that groups similar documents together. For this visualization, the cosine metric distance is used which computes distances between rows. When computing the distance between rows, it is possible to capture the relationships between sentiments as in **??**

4.4 Use case

Data A series of events that will demonstrate the application of sentiment analysis on Twitter data collected from five distinct categories is presented. The data are conducted from mainly three domains: natural disasters, sports events, and politics. Both sports events and politics are collected tweets from two different events, shown in 4.2. Each category provides insights into the challenges and opportunities associated with sentiment analysis in these specific domains, shedding light on the potential of sentiment analysis for understanding public sentiment and opinion on Twitter.

To collect the data, Twitter API v1.0 was used to access tweets from Twitter users. The API offered by the Twitter service is a moving target and has changed several times even over the course of our investigation. For access, the API has specific terms and usage policies that developers must adhere to. These policies cover areas such as data privacy, display requirements, and limitations on data usage. The most common and consistent method for

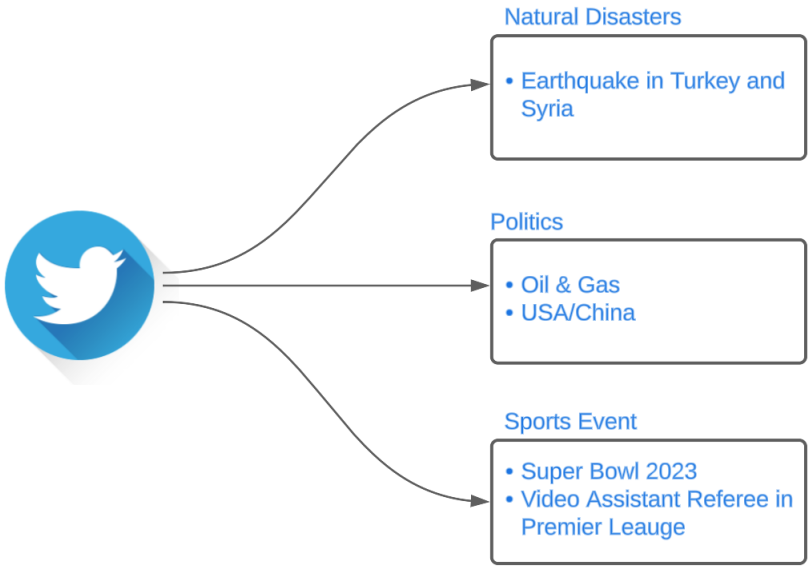

Figure 4.2: The different categories and events that are used to extract data from Twitter

gathering data is to request a paged set of data for a given query. By utilizing the API for downloading tweets the opportunity to access real-time and historical data was possible. The subjects selected for this research are keywords related to each event and are collected through queries in Python where there was entered up to five keywords for each search for Tweets. In total there where collected total amount of the 20 tweets per event. Each tweet is given to both

Natural Disasters

Natural disasters can have a profound impact on individuals and communities, leading to a surge in discussions and sentiments on social media. In this case study, we analyze tweets related to an earthquake that struck both Turkey and Syria around the time of data collection. As the event says, the expected response from people will be upset and sad. By examining the sentiment trends and patterns, we can gain insights into the level of distress, support, and resilience within the affected population, and may detect data that provides hope and tries to find some positive outcome.

The case study explores the challenges of sentiment analysis in the context of natural disasters, the collected data were manually checked to filter out duplicates and are being compared to the prompt in 4.1.

The following background information is given to the GPT-3:

On February 6, 2023, a powerful earthquake hit southeast Turkey (officially the Republic of Türkiye) and the northwest region of war-torn Syria, leaving millions of people in urgent need of basic necessities like shelter, food, clean water, and sanitation. [22]

Politics

USA & China Political events and conflicts often generate heated discussions and diverse opinions on Twitter. This research focuses on sentiment analysis in the domain of politics, specifically examining tweets related to the USA/China and at the time, a conflict over spy allegations, allegedly in the airspace of the United States.

The political domain is quite challenging to study for sentiment analysis. The use of politically charged language, identification of sentiment manipulation and propaganda, and the analysis of sentiment dynamics during ongoing conflicts contribute to making it extra difficult. In advance, this category is seen as the category where there will be the greatest chance of bias, due to the origin of Twitter and that the data will not contain the opinions of both sides of the conflict.

The following background information is given to the GPT-3:

The United States said it had concluded recovery efforts off South Carolina to collect sensors and other debris from a suspected Chinese surveillance balloon shot down by a U.S. fighter jet. [1]

Oil & Gas The global politics of oil and gas supply and demand involve complex negotiations, conflicts, and power dynamics from several of the most powerful countries in the world. With this data, we wanted to explore another field of big politics that people expressed themselves about.

The following background information is given to the GPT-3:

In the past twelve months, the oil market has absorbed the impact of Russia's invasion of Ukraine and the sanctions imposed in response by the United States, the European Union, and their allies in Asia. [9]. Russia's crude and fuel exports have been redirected to South and East Asia, while former markets in Europe have been backfilled with crude and products from the Middle East and Asia.

Sports event

Super Bowl Final 2023 Sports events, particularly high-profile tournaments like the Super Bowl, attract significant attention and generate a wealth of conversations on Twitter. By analyzing tweets from what could be fans, sports analysts, and media outlets, we can uncover

the emotions associated with the game and identify how people for example reach the winner.

Through this analysis, we aim to showcase the importance of context and domain understanding in sentiment analysis on Twitter during sports events.

The following background information is given to the GPT-3:

The Kansas City Chiefs beat the Philadelphia Eagles 38-35 on Sunday to win their third Super Bowl championship in franchise history. [21]

Video Assistant Referee in Premier League The introduction of Video Assistant Referee (VAR) in the Premier League has sparked intense debates and discussions among football fans. This topic is not new when collecting the data, but by exploring sentiment analysis in the context of VAR controversies to the public because there are new situations towards this technology in every game played. By analyzing tweets related to VAR decisions, referee controversies, and fan reactions, we can gain insights into the acceptance, criticism, and overall sentiment surrounding the use of VAR in football, in contrast to the other sports category.

The following background information is given to the GPT-3:

Premier League referees are being instructed by the PGMOL, the body responsible for elite match officials in the English game, to make more use of pitchside monitors. [5].

When selecting the following background information to analyzing sentiment in diverse domains, we can gain a holistic understanding of how emotions are expressed and perceived across different topics. This broadens the scope of the research and allows for a comprehensive analysis of sentiment in various contexts, contributing to a more understanding of how models detect and present emotions and reactions. At the same time, this research are dealing with a small amount of manually analyzed data. It is therefore important to have in mind that the framework is developed of using already existing technology in a new way for a more trustworthy and perhaps reliable framework.

4.5 Evaluation of RQ1: To what extent does the inclusion of contextual information influence the results of LLMs in sentimental analysis tasks?

In this research project, the proposed conceptual framework aimed at investigating the impact of including contextual information on the results of sentimental analysis tasks conducted using large language models (LLMs). Our primary focus is to explore the extent to which the inclusion of contextual information influences the outcomes of sentiment analy-

ID	Tweet	Event	Emotion without Context	Emotion with Context
1	"Syria is in dire need of help, but the world's focus seems to be more on Turkey."	Natural Disaster	Need, desperation	Concern, disparity
2	"Qatar is sending 10,000 container homes for earthquake victims in Syria and Turkey. These were used as World Cup accommodations."	Natural Disaster	Generosity, concern	Gratitude, appreciation, relief, hope, humanitarianism
3	"Chinese cars are now purchased in Russia, the era of gas and oil is also over. Europe is closed, we've been thrown out of it - Russian propagandist Mardan is sad."	Politics	Sad	Resignation, disappointment, sarcasm
4	"Labour are currently opposed to new oil & gas licences in the North Sea, fracking and coal mining. Their pursuit of ridiculous #NetZero goals are worse than the current plans from Sunak's Clown Cabinet."	Politics	Angry	Frustration, criticism
5	"I wanted Philadelphia to win, it was a fantastic game, the call was tough but technically a good one, the better team won. Philadelphia couldn't put them away, Kansas City dominated the second half, and the TD fumble return impacted the outcome far more than one call."	Sports event	Positive	Disappointment

Table 4.4: Results from a selection of sentiment analysis of both approaches

sis tasks.

The framework consists of several key components and steps. The research questions serves as our guiding principle throughout the research process. To implement the framework, we gather a diverse set of textual data, from social media posts and news articles for the context part. The different categories are all designed and given prompts based on the context that could provide facts about the event and contains facts, and not information the model can detect as emotion. The prompt for all events is the same, but the section where the news article/context is provided is unique for each event. As mentioned earlier, the important part was to give related context that described the current topics at the time, based on factual information. To answer the research questions,RQ1 1.1 will be discussed in the next section.

The tweets in 4.4 are selected with the aim to show and highlight potential differences in sentiment analysis results. The chosen tweets are curated to present varying contexts and expressions, allowing for an exploration of the disparities that may arise in the analysis process. As shown in 2.2.4, both approaches tend to describe events with the same sentiment group but use different words for it. The first line in 4.4, "Syria is in dire need of help, but the world's focus seems to be more on Turkey" expresses a sentiment of concern and highlights a sense of disparity. The tweet suggests that Syria requires urgent assistance, emphasizing the gravity of the situation. At the same time, the fact that people are in need and it is a desperate situation is not wrong. The fact

When analyzing the sentiment with context, the identified emotions are concern and disparity. The sentiment of concern can reflect the tweet's emphasis on the urgent need for help in Syria. It conveys a sense of worry or unease regarding the situation. The sentiment of disparity captures the perception that there is a discrepancy in the world's attention, implying that Syria's needs are not receiving the same level of focus as Turkey's.

At the same time, the sentiment analysis of the same text, but without context identifies the emotions of need and desperation. The sentiment of need captures the tweet's emphasis on Syria's dire need for assistance. It highlights the urgency and necessity of aid. The sentiment

of desperation suggests a state of extreme urgency or despair, indicating the severity of the situation in Syria.

Overall, the sentiment analysis with context captures the broader perspective of concern for Syria's situation and the perceived disparity in attention between Syria and Turkey. On the other hand, the sentiment analysis without context focuses more specifically on the immediate needs and desperation expressed in the tweet. Both analyses provide valuable insights into the sentiment expressed in the text, offering different layers of understanding depending on the inclusion or exclusion of contextual information.

In 4.4, the tweet on the second line, "Qatar is sending 10,000 container homes for earthquake victims in Syria and Turkey. These were used as World Cup accommodations.", gives a bit more different response. Without considering the surrounding context, the sentiment analysis classified the tweet as having a positive sentiment. The emotions detected in this analysis were "generosity" and "concern." This suggests that the tweet expresses a sense of "positive" sentiment, indicating acts of generosity and concern toward the earthquake victims in Syria and Turkey. However, without considering the broader context, we could say that this analysis provides an understanding of the emotions conveyed in the tweet. When a human reads the tweet, you can sense a positive action in an otherwise bad situation.

On the other side, when taking into account the contextual information provided for the tweet, you could say that a more nuanced sentiment analysis was conducted. The sentiment and emotions detected include "gratitude," "appreciation," "relief," "hope," and "humanitarianism." In the matter of analyzing this response manually as done here, the emotions that are being reflected could be a more comprehensive understanding of the tweet's content. The presence of emotions like gratitude and appreciation suggests that the tweet expresses thankfulness towards Qatar for sending container homes to aid earthquake victims. The emotions of relief and hope indicate a positive outlook and optimism for the affected regions, while the presence of humanitarianism reflects the underlying compassionate and altruistic intent behind the actions described in the tweet.

At the same time, despite using different words for describing the data, both tend to find positively related sentences based on a more negatively loaded event.

When trying to answer RQ1 1.1, the results above show that the interpretation of sentiment in tweets can vary depending on the individual human reader. It is difficult to come up with a concrete solution because you relate to feelings and thoughts that are completely different from person to person. Just like the models, the readers may arrive at different conclusions regarding the sentiments expressed in a tweet based on their own subjective understanding and biases. Human readers bring their own context, knowledge, and understanding to the analysis process. They can pick up on subtle cues, sarcasm, or irony that may be challeng-

ing for automated models to detect accurately. Additionally, humans possess the ability to consider the broader context, draw on prior knowledge, and make more nuanced judgments based on their understanding of the specific topic or event. When it can be so difficult for humans to detect a correct response, it is safe to say that it is hard for a computer.

Nevertheless, while sentiment analysis models can provide automated assessments of sentiment, they may not always capture the full complexity and nuances of human emotions and intentions. Contextual information, such as the broader conversation, user profiles, or the specific event being discussed, can significantly influence how a tweet is perceived and interpreted.

Ultimately, the inclusion of human readers and their subjective interpretations is crucial in evaluating sentiment analysis results. By considering the human factor, we can gain a more comprehensive understanding of the sentiments expressed in tweets and ensure that sentiment analysis approaches align with human perceptions and expectations.

With sentiment keywords describing the event, it is possible to extract patterns in where usual sentiment each category/event is common for. It is possible to see two cases: With context, emotion detection can be more trustworthy,

When detecting emotions like this, you will have the possibility to look at both positive and negative outcomes from the Twitter event.

When analyzing tweets and performing sentiment analysis, including context as a prompt often leads to better sentence generation that accurately describes the actual tweet. However, it is important to note that this is not universally applicable to every event or happening.

Incorporating context as a prompt provides additional information and background that aids in understanding the intended meaning and emotional tone of the tweet. By including relevant contextual cues such as preceding dialogue, subject matter, or the specific event or situation being discussed, the model gains a frame of reference for interpreting the tweet's sentiment.

The presence of context allows the model to disambiguate the sentiment expressed in the tweet and generate sentences that are more coherent and contextually appropriate. With a broader understanding of the context, the model can capture the nuanced aspects of sentiment and generate more accurate responses.

It is important to acknowledge that the effectiveness of using context as a prompt may vary depending on the event or happening being analyzed. In some cases, the provided context may not significantly impact the quality of sentence generation, and the model's performance may be more reliant on the inherent capabilities of the sentiment analysis algorithm.

Certain events or happenings may have sentiments that are more directly conveyed in the tweet text itself, without requiring additional context for accurate interpretation. In such cases, the inclusion of context as a prompt may not yield substantial improvements in sentence generation.

Therefore, while incorporating context as a prompt has been shown to enhance sentence generation in sentiment analysis, it is essential to consider the specific event or happening under analysis. Careful evaluation is necessary to determine the extent to which context influences the accuracy and relevance of the generated sentences.

In conclusion, incorporating context as a prompt in sentiment analysis can often lead to improved sentence generation that more accurately describes the tweet. The impact of context may vary depending on the event or happening being analyzed, but a human, who does not have knowledge of the processes behind the analysis itself, is given a false sense of security when providing sentiment analysis with prompts containing context information. It is far more important to be aware of the degree of difficulty in blindly trusting processes and analyses that are themselves difficult for humans. Therefore, context on a case-by-case basis could ensure the validity and reliability of sentiment analysis results.

4.6 Results from Survey

To both answer the research questions and give a wider analysis of how context impacts the sentiment analysis, a survey is created. The sentiment analysis survey is conducted with 50 participants, mainly from the United Kingdom, but also with participants from the United States, and China. The purpose was to achieve two primary objectives while considering diverse cultural perspectives. Firstly, the survey sought to determine whether participants would choose contextually relevant sentences from the provided options. Secondly, participants were asked to rate the perceived suitability of the chosen sentences on a scale from 1 to 5.

This survey's dual purpose of evaluating contextual sentence selection and capturing participants' perception of fit on a numerical scale contributes to the advancement of sentiment analysis techniques. By considering the cultural diversity of the participants, the survey findings offer valuable insights into cross-cultural differences and similarities in sentiment interpretation. These insights enhance our understanding of sentiment analysis within diverse cultural contexts and facilitate the development of more robust and culturally sensitive sentiment analysis models.

The selection of questions for the survey was driven by an intriguing observation: many orig-

Number	Tweet	Event	Emotions without Context	Emotions with Context	Neutral
1	" Syria is in dire need of help, but the world's focus seems to be more on Turkey."	Natural Disaster	7	32	11
2	"haechan donated 100 million won to help victims of turkey and syria recover from the recent earthquake"	Natural disaster	11	13	26
3	"Qatar is sending 10,000 container homes for earthquake victims in Syria and Turkey. These were used as World Cup accommodations."	Natural disaster	11	12	27
4	"Anonymous donor gives stunning $30 Million toward Turkey/Syria earthquake victims"	Natural disaster	19	11	20
5	"FUNDRAISER! Hello all, we made this fundraiser for a dear Kurdish friend who lost their home and belongings during the earthquake. Please boost! All help is appreciated."	Natural disaster	18	9	23
6	"'Amin Nasser, chief executive of Saudi Aramco, said mounting pressure to curb new investment in oil and gas was based on flawed assumptions...'"	Politics	5	39	6
7	"BREAKING NEWS SAUDI ARAMCO IS WARNING THAT AN INCREASED FOCUS ON CLIMATE WAS UNDERMINING INVESTMENT IN OIL AND GAS TO THE POINT WHERE IT NOW POSED A THREAT TO THE WORLD'S ENERGY SECURITY This is one of the biggest companies in the world. The narrative is changing..."	Politics	5	35	10
8	"Chinese cars are now purchased in Russia, the era of gas and oil is also over. Europe is closed, we've been thrown out of it - Russian propagandist Mardan is sad."	Politics	18	15	17
9	"it's the middle of winter. syrians are already suffering from inhumane sanctions. 2-3 hours of electricity a day, no gas, no oil, no hot water, no wood for fire places. the cost of living has become unbearable and people barely have enough to eat. this earthquake has broken them."	Politics	11	10	29
10	"Canada confirms that oil and gas activity, mining, the dumping of certain waste materials and destructive bottom trawling fishing won't be allowed in any Marine Protected Areas"	Politics	17	18	15
11	"Labour are currently opposed to new oil & gas licences in the North Sea, fracking and coal mining. Their pursuit of ridiculous #NetZero goals are worse than the current plans from Sunak's Clown Cabinet."	Politics	15	17	18
12	"Philadelphia Eagles fans flipped over a vehicle before their Super Bowl showdown against the Kansas City Chiefs last night.The Chiefs secured a 38-35 victory over the Eagles in the final minutes of the game. "	Sports	12	36	2
13	"I wanted Philadelphia to win, it was a fantastic game, the call was tough but technically a good one, the better team won. Philadelphia couldn't put them away, Kansas City dominated the second half, and the TD fumble return impacted the outcome far more than one call."	Sports	2	46	2
14	"What high schools were represented in Super Bowl I? Here are the rosters for the Green Bay Packers and Kansas City Chiefs along with each player's high school"	Sports	2	46	2
15	"Prince Tega moved to the United States from Nigeria at the age of 16yrs old - with two pairs of boxers and just $20 in his pocket. 9yrs later, he won the Super Bowl with Kansas City Chiefs."	Sports	6	28	16
			159	367	224

Table 4.5: Evaluation results from the survey

inal responses, both with and without context, generated similar sentiment analysis results when processed by the automated sentiment analysis model. However, a subset of questions with varying results captured the attention and these specific questions were included in the survey. Combined with the more similar responses, they presented interesting scenarios where human judgment could potentially diverge from the model's analysis, or give a stronger picture of the assumptions.

The inclusion of such questions offers an opportunity for participants to exercise their human judgment and contribute their individual perspectives without relevant information to which text is created from which prompt. These variations in responses between the model and human participants provide valuable insights into the limitations of automated sentiment analysis and highlight the need for human involvement in certain nuanced situations.

The tweet presented in 4.3 is the same as in 4.5. "Syria is in dire need of help, but the world's focus seems to be more on Turkey.".

The first line refers to the non-contextual sentiment analysis and is: "The text contains feelings of need and desperation." The second line is showing the contextual-based approach with the result "The text expresses concern about the lack of attention given to Syria despite their dire need for help, with the world's focus seemingly more on Turkey after the earth-

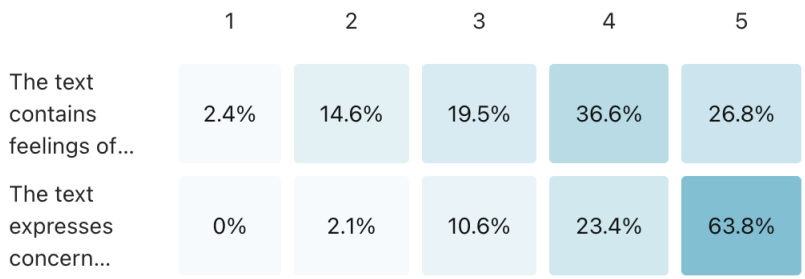

Figure 4.3: A question from the survey

quake"

63.8% of the responses show a clear majority who prefer the text that actually describes both the sentiment and the text they are presented for. Participants' responses demonstrated variations in sentiment perception depending on the specific context in which the text was presented. It shows that the more context that is provided, the easier it is to choose that answer. The mentioned context sentiment was chosen from 33 of 47 responses as the best suit.

As expected before conducting this research, the output could vary from event to event. The variation in output from event to event in sentiment analysis is expected, but in **??**, the events related to Natural Disasters are mostly neutral selected from the participants. If that is because the responses look similar, and time pressure or other factors mean that the safest choice is, in principle, not to choose is difficult to know. Next, the participants are presented with six tweets in the domain of politics. The findings indicate that participants tend to prefer to choose the text with contextual analysis. The notable difference from the table presented in answering the first research question is that the participants are provided with sentences, not just keywords. The inclusion of contextual information can provide valuable insights and enhance the overall understanding and interpretation of sentiments expressed in the texts. An example of this can be that the participants could trust the text more because the comment they select contains parts of the text they are to analyze. Regardless, the total score shows that Sentences containing "Emotions without Context" was selected 159 times, "Emotions with Context", 367 times, and the neutral 224 times. The primary reason for the big difference lies in the Sports Events and Politic categories.

4.7 Evaluation of RQ2: To what extent does context play a role in making social media comments more trustworthy and reliable for humans?

By incorporating these sentence-based questions, the survey aimed to bridge the gap between non-context sentiment analysis and context-based more human-like perception. With the emergence of AI in mind, and its increasing availability to individuals who may not possess knowledge of its inner workings, also considering the concerns regarding the potential hallucination problem 2.2.3. The survey clearly shows that participants tend to choose the context-based approach.

At the same time, the primary difficulties of the models are to generate outputs that appear convincingly real, even though they may lack a factual basis. This can lead to misinformation, false beliefs, and the perpetuation of inaccurate or biased information.

For instance, an AI-generated text that presents itself as factual news or expert advice, but is actually fabricated or lacks proper verification, can mislead individuals who rely on it without critically assessing its validity. Therefore, some of the same conclusion as with RQ1 can be made. Incorporating context as a prompt can lead to improved sentence generation, and is often easier to evaluate when humans are doing it them self, then fully trust a model to do so. The connection between the sentences the participant is assigned and how they will be impacted is important to be aware of when using or trust these types of results. Because, the results often agree with feelings we ourselves have after reading a text, but as research has shown it is difficult to know exactly when it will be wrong. This means that there will always be a possibility that you will make your choices on the wrong premises.

Chapter 5

Summary and Recommendations for Further Work

5.1 Summary and Conclusions

The research explores the impact of context on sentiment analysis in social media comments. The goal was to investigate how the inclusion of contextual information affects the trustworthiness, and reliability of sentiment analysis outcomes. The research employed a design science research approach and used a model-driven framework for healthcare information visualization as a basis for analysis.

The study found that incorporating contextual information in sentiment analysis can lead to more trustworthy results, but it is not possible to say for sure as we're discussing human aspects that are not measurable with numbers or conclusions. Contextual cues provide additional insights and help in understanding the true sentiment expressed in social media comments. The analysis demonstrated that when the context was provided as a prompt, the generated sentences describing the sentiment of the tweets aligned more closely with the actual content, but at the same time sentences provided without context were often responding with information from the input data. However, it was also noted that the impact of context varied across different events and happenings. While context generally improved the quality of sentiment analysis, it was not universally effective in all cases. The complexity of certain events or the absence of explicit emotional words in tweets posed challenges in accurately capturing sentiment.

The research emphasized the importance of correct contextual information in sentiment analysis, and that one must be aware of these challenges when using this type of model. Context can definitely give trustworthy results, enhancing their usability in real-world applica-

tions. The research highlighted some limitations of sentiment analysis, such as the potential for the model to make simple mistakes. These types of limitations call for further advancements in natural language processing models to improve their accuracy and robustness in sentiment analysis tasks.

5.2 Recommendations for Further Work

Determining how to provide context in such scenarios requires careful consideration. One potential approach is to select a subset of relevant articles that encapsulate the necessary context. However, this raises the question of which articles to include. Should the latest article be prioritized? While considering the latest article may provide a more up-to-date context, it has its drawbacks.

Future work in this domain could involve exploring automatic systems for sentiment detection in tweets that are not common to a specific event or field. It could also investigate the role of different data sources, such as Reuters, in providing reliable context for sentiment analysis. The research could benefit from integrating advanced techniques. Twitter is known for its real-time nature, where tweets are posted and shared rapidly. Future research should explore techniques for real-time sentiment analysis on Twitter, enabling businesses, journalists, or organizations to monitor and respond to sentiment trends in a timely manner. Efficient stream processing, incremental learning, and online adaptation techniques can be explored to facilitate real-time sentiment analysis on Twitter.

Overall, this research contributes to the understanding of the influence of contextual information on sentiment analysis in a zero-shot approach. t can be interesting to further be able to analyze the reason why it stands out from the rest and light on the importance of context in accurately capturing sentiment and highlights the potential of using context-aware approaches to enhance the reliability and trustworthiness of sentiment analysis outcomes. Further investigation into how the increasing use of humans without knowledge not could fully trust models and their output.

References

[1] Ali, I. and P. Stewart (2023, February). Chinese spy balloon flies over the united states: Pentagon. https://www.reuters.com/world/suspected-chinese-spy-balloon-flying-over-united-states-us-officials-2023-02-02/.

[2] Center, P. R. (2019, April 24). Sizing up twitter users: Methodology.

[3] Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

[4] Global Industry Analysts, I. (2023). Sentiment analytics - global strategic business report 2023. *ResearchandMarkets*. Accessed on May 11, 2023.

[5] Herman, M. (2020, January). Soccer-premier league refs told to use pitchside monitors for red cards. https://www.reuters.com/article/soccer-england-var-idINL8N29M4GO.

[6] Hevner, A. and S. Chatterjee (2010). *Design Research in Information Systems: Theory and Practice*. Springer.

[7] Hevner, A. R., S. T. March, J. Park, and S. Ram (2004). Design science in information systems research. *MIS Quarterly 28*(1), 75–105.

[8] Hu, X., B. Zhang, C. Lu, and Q. Li (2013). Unsupervised sentiment analysis with emotional signals. *IEEE Intelligent Systems 28*(2), 76–80.

[9] Kemp, J. (2023, February). Oil market has fully absorbed impact of russia's invasion of ukraine. https://www.reuters.com/lifestyle/sports/kansas-city-chiefs-beat-philadelphia-eagles-win-super-bowl-2023-02-13/.

[10] Kiritchenko, S., X. Zhu, and S. M. Mohammad (2014, August). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research 50*, 723–762.

[11] Li, B. and P. Lu (2020). Convert word to vector component. [Online; posted 11/04/2021].

[12] Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

[13] Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 142–150. Association for Computational Linguistics.

[14] Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space.

[15] Narayanasamy, S. K., K. Srinivasan, S. Mian Qaisar, and C.-Y. Chang (2021). Ontology-enabled emotional sentiment analysis on covid-19 pandemic-related twitter streams. *Frontiers in Public Health 9*.

[16] OpenAI (2021). Gpt-3.5 model documentation. https://platform.openai.com/docs/models/gpt-3-5. Accessed on May 3, 2023.

[17] Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe (2022). Training language models to follow instructions with human feedback.

[18] Pak, A. and P. Paroubek (2010, 01). Twitter as a corpus for sentiment analysis and opinion mining. Volume 10.

[19] Pang, B., L. Lee, and S. Vaithyanathan (2002, July). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86. Association for Computational Linguistics.

[20] Radford, A. and K. Narasimhan (2018). Improving language understanding by generative pre-training.

[21] Reuters (2023, March). Kansas city chiefs beat philadelphia eagles to win super bowl. https://www.reuters.com/business/energy/oil-market-has-fully-absorbed-impact-russias-invasion-ukraine-kemp-2023-03-09/.

[22] Reuters (2023). Pictures of the month: February. https://www.reuters.com/news/picture/pictures-of-the-month-february-idUSRTSGPT7N. Accessed: 13 March 2023.

[23] Schick, T. and H. Schütze (2020). Exploiting cloze questions for few-shot text classification and natural language inference. *CoRR abs/2001.07676*.

[24] Serrano, L. (2023, April). What are transformer models and how do they work? [Online; posted 12-April-2023].

[25] Shaw, P., J. Uszkoreit, and A. Vaswani (2018). Self-attention with relative position representations. *CoRR abs/1803.02155*.

[26] Singh, I. (2020, July). Syntactic processing for nlp. [Online; posted 1-July-2020].

[27] Tang, D., F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin (2014, June). Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 1555–1565. Association for Computational Linguistics.

[28] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is all you need. *CoRR abs/1706.03762*.

[29] Wang, Z., V. Joo, C. Tong, X. Xin, and H. C. Chin (2014). Anomaly detection through enhanced sentiment analysis on social media data. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pp. 917–922.