MASTER'S THESIS PROPOSAL

Developing a Sentiment Analysis Model for German Social Media Data Using NLP Techniques

NOVEMBER 8, 2024

GISMA BUSINESS SCHOOL
AI, Data Science and Digital Business

Contents

1. Introduction

The swift development of social media in Germany has translated into an enormous amount of user-generated content, thus providing the chance to discover public opinion on issues such as consumer preferences and political positions. For businesses, government officials, and research institutions, this sentiment evaluation is essential to comprehend the public climate and to align with the direction of society.

Though there have been considerable improvements in Natural Language Processing (NLP), the majority of sentiment analysis models are English-based. Consequently, they are not sensitive enough for German content. The aim of the thesis is to resolve this issue by creating a sentiment analysis model particularly for German social media which will employ NLP methods to achieve better accuracy and use.

Objective The core goal of this study is to develop and implement a sentiment analysis model that will classify sentiment within German-language social media posts correctly and thus will be a basis for the development of real-time sentiment assessment tools.

2. Problem Statement

The issue is that the German language is a complicated one because of features like compound words, grammatical structures, and colloquialisms, which makes it difficult to develop a precise sentiment analysis model for it. Even though there are many sentiment analysis tools, they mostly fail to recognize sentiment in the German-language posts due to their restricted capabilities, thus they are not that useful for German audiences and businesses. This thesis focuses on the need for a powerful model that could understand the specific features of the German text and produce an accurate analysis of the sentiment.

3. Research Questions

What preprocessing and NLP techniques are applied to German-language social media that are more accurate in sentiment analysis?

Which are the existing NLP models that we can adapt to the German context to better understand the sentiment?

How accurate and well-performing can state-of-the-art models that are based on transformer-based architectures achieve for German social media datasets?

4. Literature Review

Many investigations have been done for sentiment analysis in English social media, applying machine learning and NLP techniques like LSTM (Long Short-Term Memory networks) and transformer models (e.g., BERT). With regard to German, though, research is still scarce, and there are few models dealing with German nuances. The literature review will be centered on the latest developments in NLP, particularly such models as multilingual BERT, which have been used to transform into other languages besides English and thus adapt to German language data.

5. Methodology

5.1 Data Collection

To generate a dataset of superior quality, data will be extracted from those platforms that German users actively use to discuss current issues, products, or trends. Twitter and Reddit are two possible sources since they provide APIs and most of the content is in German.

Data Volume: We expect to compile a database of 50,000-100,000 posts for training and testing of the model.

Ethical Considerations: Data collection will conform to the platform usage rules, making the information of users anonymous and ensuring that the data is used under the privacy laws.

tokens and lemmatization are some of the most important steps in the process.

5.2 Data Preprocessing

The role of data preprocessing is to enhance model performance. The preprocessing steps are Language Detection: Including only German-language posts.

Tokenization and Lemmatization: The process of turning text into a format that is suitable for analysis, while at the same time, managing German-specific problems such as compound words.

Labeling: The semi-supervised method might be used for initial labeling and manual checks will be done to confirm the correctness of sentiments.

5.3 Model Development

The transformer architectural models, which are the basis of the proposed approach will be considered as successful ones in NLP. The German BERT and its German-specific version, German BERT will be assessed for this purpose. The model's training will consist of:

Feature Engineering: Developing emotion-based features that express the emotional tone.

Model Training: The model will be supervised through labeled data to learn how to classify sentiments.

Evaluation Metrics: The key measures of the model's performance are accuracy, precision, recall, and the F1-score.

6. Expected Outcomes

This research seeks to establish a sentiment analysis model that is highly accurate and reliable for German social media texts. The expected results are:

A whole sentiment analysis system made in German, tangible in such sectors as marketing, media, and public relations.

A comprehensive evaluation report detailing the model's capabilities in comparison with other solutions.

The project additionally seeks to augment the research literature on sentiment analysis in non-English languages, by presenting solutions for two problems: the lack of German data and the out-of-the-box usage of sentiment analysis.

7. Business Application

7.1 Target Market

Real-time emotion insights can be used for e-commerce, media, and government agencies to make effective decisions in the respective processes. Sentiment model that focuses on German-speaking audiences will allow them to have a clearer picture of the public's opinion apart from acting proactively.

7.2 Value Proposition

The model is developing a customized solution for German-speaking businesses and public institutions to measure their customers and the general public to make them effective. It comes with the main advantage of addressing such issues over the normal sentiment analysis tools that are being used.

7.3 Revenue Model

The project can generate revenue through:

Subscription Services: Through an arrangement of regular reporting of sentiment analysis to companies.

API Access: The capability of a company to incorporate the sentiment model into their data pipelines will be available via the interface.

8. Work Plan and Timeline

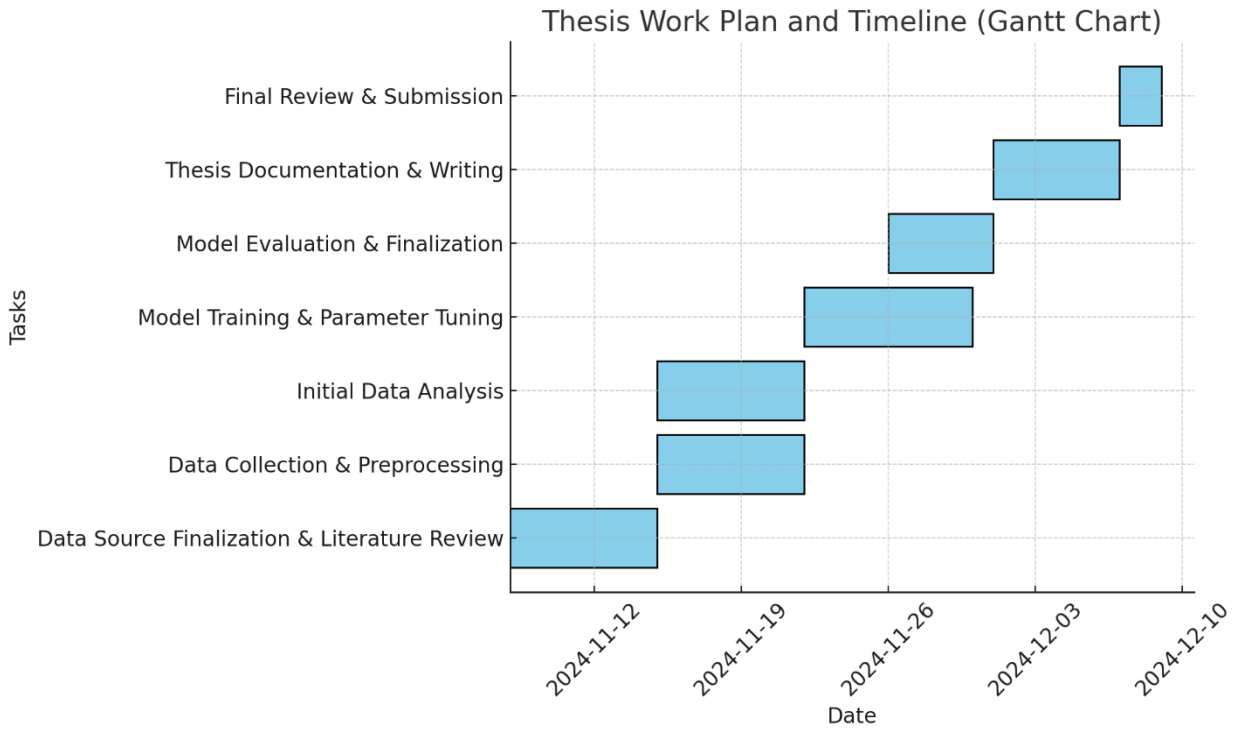
The master's thesis must be finished in a month and followed a structured weekly schedule as described below:

Week 1: Finalize data sources, include the literature review and start collecting data.

Week 2: Data processing and primary data analysis finished. To also start modeling.

Week 3: To keep model training and the process of parameter adjustment smooth. Analysis should be carried out when the model is complete, and the model should be finished.

Week 4: write a thesis paper, which includes methodology, results, and conclusions. Then, make a preparation for submission.



9. Conclusion

The study will work on a German-language-based sentiment analysis model that will use the characteristics of German language in a wider range of words. This method, due to its reliance on state-of-the-art NLP methods for real-time sentiment analysis, is promising to result in a practical tool that covers communicational challenges and potentials for German language audiences.

10. References

A preliminary list of key sources includes:

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Kübler, S., & Zinsmeister, H. (2015). Corpus Linguistics and Linguistically Annotated Corpora.

BERT (2020) German Model Documentation.