Comparative Analysis of Sentiment Analysis Techniques for Social Media Data Using Deep Learning Models

Daniel Coates

***Abstract*—This paper wanted to compare the performance and accuracy of sentiment analysis models that use generic word embeddings like GloVe and word2vec against embeddings trained on a corpus of social media data. The hypothesis is that domain-specific word embeddings improves the accuracy of sentiment analysis on social media data. This is expected due to domain-specific embeddings better capturing the nuances of language more accurately leading to better performance.**

I. Introduction

Progress in NLP models has increased exponetially in the last year. ChatGPT the most well known NLP model reached 100 million monthly users 2 months after launch and is the fastest growing conumer application in history [**?**] Sentiment analysis is the process of automatically detecting the sentiment of a piece of text. [**?**] This is a crucial part of NLP models and has gained increasing attention in recent years. Deep learning models such as recurrent neural networks (RNNs) and convoluted neural networks (CNNs) can outperform traditional machine learning models and have shown promising results. However, the quality of the word embeddings that these models are trained on directly affects the performance of these deep learning models. In this paper the performance of sentiment analysis techniques using deep learning models will be compared. In particular, the effectiveness of domain-specific word embeddings in improving sentiment analysis accuracy on social media data. The hypothesis is that domain-specific embeddings will better capture nuances in language more accurately, leading to better performance in sentiment analysis compared to generic embeddings. To test this hypothesis several deep learning models such as CNNs and RNNs will use both domain-specific and generic embeddings and be compared to each other. The performance will be based off several metrics such as accuracy, precision,recall and F1 score. The paper will be organised as follows. In section II, a literature review on sentiment analysis techniques,word embeddings and deep learning models. In section III, we will discuss the methodology of the experiments. In section IV, we present and analyse the results before discussing the implications of the experiments in section V. Finally, the paper will conclude in section VI.

II. Literature review

Sentiment analysis has been studied extensively in recent years, especially for social media data. Deep learning models have also showed promising results over traditional machine learning models such as support vector machines, due to their ability to capture more complex patterns and features in data. [**?**]

Word embeddings are an important part of sentiment analysis. The represent words as vectors in many dimensions.[**?**] Word embeddings contain syntactic and semantic information about the words.

However, the quality of word embeddings has a significant impact on the performance of these models. Generic word embeddings are trained on large amounts of data from various sources. For example, the popular word2vec embedding was developed by google and trained on the google news dataset containing roughly 100 billion words.[**?**] Domain specific word embeddings trained on data from a specific domain may better capture domain specific language, leading to better sentiment analysis.

There have been several studies looking at the use of domain specific word embeddings on social media data. For example, Twitter Sentiment Analysis with CNNs and LSTMs[**?**] investigated the performance of word embeddings on twitter data, including domain specific embeddings to detected the sentiment behind tweets. The results were promising, showing better accuracy and F1 scores from the domain specific embeddings.

In summary, prior research has demonstrated that domain-specific word embeddings can enhance the performance of deep learning models for social media sentiment analysis. In order to further assess the effect of domain-specific word embeddings on performance, we compare sentiment analysis methods employing deep learning models with both generic and domain-specific word embeddings.

III. Methodology

Firstly, data needs to be collected from various social media sites. The data will be related to to pick. The data will then be labelled with their corresponding sentiment (positive, negative, neutral).

This data will the be preprocessed to remove noise and irrelevant data, such as punctuation, URLs, hashtags and mentions. The data will also be normalised using stemming and lemmatisation, to reduce the dimensionality of the data.

We will then extract features from the preprocessed text using generic word embeddings such as choose embedding. For the domain specific embedding I will train my own word embedding using a large corpus of text from the social media sites.

We will then train both CNNs and RNNs on both the generic and domain specific word embeddings and evaluate their performance. To evaluate performance several metrics will be used, for example, accuracy, precision, recall, and F1 score.
Finally, we will compare the performance of different models and input features and interpret the results to draw conclusions about the impact of domain-specific word embeddings on sentiment analysis performance.