CAIRO UNIVERSITY

Twitter Sentiment

ABSTRACT:

Social media has become essential at everyday life, everyone has a social media account nowadays.Twitter (X) has become one of the largest social media platforms with millions of people on them everyday.People expressing their opinions everyday and commenting on the platform about certain products or situations. Sentiment analysis which is a sub-field of NLP text mining have become a widely popular field because its capability to analyse comments and tweets of people and analyze their opinions,interests and use them to predict recent trends,stocks and help brands manage their reputation or even political elections. Twitter Sentiment Analysis can classify tweets into categories of positive and negative opinions about a situation which makes Twitter Sentiment Analysis very useful and easily applicable for many various purposes.

KEYWORDS: sentiment analysis; text classification; natural language processing; Twitter .

1 INTRODUCTION

Social media has expanded exponentially in the past years millions of users everyday use social media to connect with the world and other people and which has produced a huge amount of user generated data like comments,tweets which express peoples' opinions and feelings about a certain subject or object of interest. Social media platforms have connected businesses to their customers,businesses can either advertise their products to customers or speak directly to them also By actively monitoring and interacting with customers on social media, businesses can gather insights, address their concerns, and build stronger customer relationships. This proactive approach to customer service can help increase customer loyalty and positive brand perception.

.Also customers can express their opinion on a specific company or a product freely.

Twitter is one of the largest social media platforms in the world with 200+ millions of users active on a daily basis.Therefore, using Twitter data for sentiment analysis has become a popular trend. The growing interest in social media analytics has focused more attention on the natural language processing (NLP) and artificial intelligence (AI) technologies involved in analytics.

Sentiment analysis, also known as opinion mining, is a branch of natural language processing that involves the use of computational techniques to determine the sentiment or emotional tone expressed in a message. paragraph. It aims to understand and classify whether the sentiment expressed is positive, negative or neutral.

The process of sentiment analysis involves analyzing various text elements, such as words, phrases, or even emojis, to evaluate the underlying sentiment conveyed by the author. Advanced algorithms and machine learning models are used to extract semantic meaning, context, and nuance from text data.

Sentiment analysis finds applications in many fields. In business, it helps understand customer opinions, feedback and preferences, thereby helping companies make data-driven decisions to develop products, marketing strategies and improve customer service . It also facilitates brand monitoring and reputation management by tracking public opinion towards a brand or product.

Social media platforms play a significant role in sentiment analysis, as they provide a vast amount of usergenerated content that can be analyzed for sentiment. By monitoring and analyzing social media discussions, businesses can gain real-time insights into public sentiment about their brand, products, or industry. Sentiment analysis also has applications in customer service, where it can be used to automate the process of categorizing and prioritizing customer feedback, identifying potential issues or complaints, and assisting in generating appropriate responses or actions. Academically, sentiment analysis contributes to research in areas such as socio-linguistics, psychology, and political science. It allows researchers to analyze public opinion on social and political issues, track sentiment trends over time, and understand the impact of events or policies on public sentiment. As sentiment analysis techniques continue to develop, there is growing interest in integrating domain-specific knowledge, contextual understanding, and multilingual capabilities into sentiment analysis model, allowing for more accurate and nuanced emotion classification. In summary, sentiment analysis is a powerful tool that leverages machine learning and language processing techniques to analyze and classify emotions expressed in text data. By providing insights into public opinion, sentiment analysis helps businesses, researchers and individuals make informed decisions, monitor brand reputation and understand popular sentiment trends .

Twitter API offers a chance to organizations to collect these data and analyse them using NLP(natural language processing) and AI(artificial intelligence) techniques to make predictions about recent trends, identify users opinion on specific products, Brand advertisement monitoring or even election campaigns monitoring and it is really easy to use! which makes Twitter a really good source for data and a treasure trove for organizations that aim to improve their profits using Sentiment analysis on twitter data.

According to a recent study shows that 87 percent of people's decision making and purchase are influenced by customer reviews. Companies use Sentiment analysis to analyse the trend and are able to respond to their customers make much more profits than the ones that doesn't. Sentiment Analysis is the process of analyzing user-generated text and determining the emotional tone from the text then categorizing it to Positive, Negative, Neutral using Natural language processing techniques. Overall, The combination of social media, user-generated data and sentiment analysis provides businesses with a powerful tool to understand customer preferences, adapt their strategies and deliver products and services that match customer expectations.

2 Related Work

When applying sentiment analysis, an important step is to classify the extracted data into different sentiment polarities, which typically include positive, neutral, and negative classes. A wide range of emotions can also be considered, which is the focus of the emerging fields of emotional computing and sentiment analysis (Cambria 2016). There are different ways of separating emotions according to different research topics, for example in political debates, emotions can be divided into satisfaction and anger (D'Andrea et al.2015). Emotion analysis with ambient emotion management can be integrated to produce more accurate results and describe emotions according to detailed categories such as
anxiety, sadness, anger, excitement and happiness (Wang et al. 2015, 2020). Sentiment analysis on Twitter has gained attention in recent years due to Twitter's widespread use as a platform for expressing opinions and sentiments. Many studies have been conducted to explore and analyze emotional patterns in tweets using different techniques and methods. In this section, we present a review of the main findings and approaches in the field of sentiment analysis on Twitter. There are three main methods for detecting and classifying emotions expressed in text: vocabulary-based approaches, machine learning, and hybrid techniques. The vocabulary-based approach uses word polarity, while the machine learning approach treats text as a classification problem and can be divided into unsupervised, semi-supervised and supervised learning.

Current SA methods can be divided into three categories: ML-based, dictionary-based, and DL-based SA.

2.1 ML-based approaches

typically use a bag of words to convert text into features. Then, the features obtained from complex ML methods are fed into classifiers such as Naive Bayes (NB), Decision Trees (DT), and Support Vector Machines (SVM) .

2.2 Dictionary-based approaches

typically collect positive and negative sentiment words in a given text to calculate the text's polarity based on the total number of these words. Unlike dictionary-based approaches, ML-based approaches can benefit from sentiment dictionaries, which consist of a series of positive and negative values assigned to different words. In this regard, ML-based approaches offer various advantages over dictionary-based approaches. In the existing literature, ML-based and hybrid dictionary-based methods have been used together. ML-based approaches were later replaced by DL-based methods, whose experimental results appeared to be more promising than those of other approaches.

2.3 Deep learning (DL)

DL methods have gained significant popularity in sentiment analysis (SA) research due to their notable achievements. For instance, Chen et al. introduced a singledimensional convolutional neural network (CNN) model that incorporated temporal relations into user and product representations. This enhancement aimed to enhance SA performance at the document level. Similarly, Liu et al. presented an artificial neural network-based approach for recommending idioms in essay writing. Their model assessed the similarity between the provided context and potential idioms. Kalchbrenner proposed a CNN model capable of analyzing introductory sentences of varying lengths. In a different vein, Tai et al. proposed long shortterm memory (LSTM) with feedback features, which improved upon the existing recurrent neural network (RNN) architecture. Furthermore, Schuster and Paliwal introduced a

bidirectional LSTM (Bi-LSTM) model that employed two distinct LSTM networks with enhanced features.

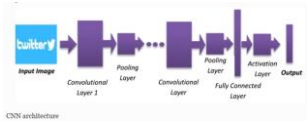
- Chen et al. proposed a single-dimensional convolutional neural network (CNN) model for sentiment analysis at the document level. Their model embedded temporal relations into user and product representations, resulting in improved sentiment analysis performance. By leveraging the hierarchical structure of CNNs, the model effectively captured local contextual information and learned discriminative features for sentiment classification.

- Kalchbrenner proposed a CNN model specifically designed to analyze introductory sentences of varying lengths. By applying convolutional operations over the input text, the model captured meaningful local patterns and learned informative representations for sentiment analysis. This allowed the model to handle different sentence structures and improve sentiment classification accuracy .

- Liu et al. presented an artificial neural networkbased approach for recommending idioms in essay writing. Their model employed similarity calculations between the given context and candidate idioms. By utilizing the power of neural networks, the model captured semantic relationships and contextual cues, enabling accurate idiom recommendations

2.4 The Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) is a deep learning approach that differs from artificial neural networks by employing a specialized architecture with distinct layers dedicated to feature extraction. This figure illustrates a typical CNN architecture, commonly utilized for image classification tasks. Inspired by the workings of the human brain, CNNs process input images through their architecture. The CNN architecture comprises convolutional layers, pooling layers, a fully
connected layer, and an activation layer. The initial layer of a CNN is the convolutional layer, responsible for extracting local features from the input image. The pooling layer, connected to a fully connected layer, is structured according to the CNN's architectural design. During the learning process, CNNs employ back propagation steps to minimize losses. Additionally, activation functions such as Softmax and Tanh are utilized to produce output values, as depicted in the figure .



CNN architecture

Overall, the application of deep learning models in Twitter sentiment analysis has shown significant promise. Researchers have explored various architectures, including CNNs, LSTMs, and their variants, to effectively capture sentiment patterns and contextual information in tweets. These models have demonstrated their ability to handle short and informal text, extract meaningful features, and achieve state-of-the-art performance in sentiment classification

3 Data set Description

Object name: Sentiment140 data set with 1.6 million tweets
Creation dates: The data set was created in 2009. Also, the data set was collected and annotated up until September 2021.
Data set Source/Creators: The data set was created by Stanford University researchers.
Repository name: The Sentiment140 data set is available on the Stanford University website.
Data Collection Process: The data was automatically created, as opposed to having human's manual annotate tweets. assume that any tweet with positive emoticons, like :), were positive, and tweets with negative emotions, like :(, were negative by using the Twitter Search API to collect these tweets using keyword search.
Language: The data set is primarily in English.

Format of the training data

It is a collection of tweets along with associated sentiment labels. The data set was created by Stanford University and contains over 1.6 million tweets that are labeled as either positive, negative, in terms of sentiment. Emojis are often used in the dataset to denote sentiment, which can provide valuable context for understanding the emotional tone of the tweets.

The data is a CSV and data file format has 6 fields/features:

0- The sentiment polarity of a tweet is represented by numerical values, where 0 indicates a negative sentiment, and 4 signifies a positive sentiment.

- the id of the tweet.

- the date of the tweet.

- if there is no query associated with the tweet, the value for this field is labeled as "NO-QUERY."

- the user that tweeted.
- the text of the tweet.

A visual representation of the data for better understanding and reference. The following screenshot displays the random columns of the dataset, highlighting key information that will serve as guidance for the analysis.

And, The forthcoming screenshot will showcase the result of applying the ”info” method to the dataset, providing valuable insights into the data types, non-null counts, and memory usage.

```
Train_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1600000 entries, 0 to 1599999
Data columns (total 6 columns):
 #   Column  Non-Null Count    Dtype
---  ------  --------------    -----
 0   0       1600000 non-null  int64
 1   1       1600000 non-null  int64
 2   2       1600000 non-null  object
 3   3       1600000 non-null  object
 4   4       1600000 non-null  object
 5   5       1600000 non-null  object
dtypes: int64(2), object(4)
memory usage: 73.2+ MB
```

3.1 Data set Preprocessing

3.1.1 Feature selection

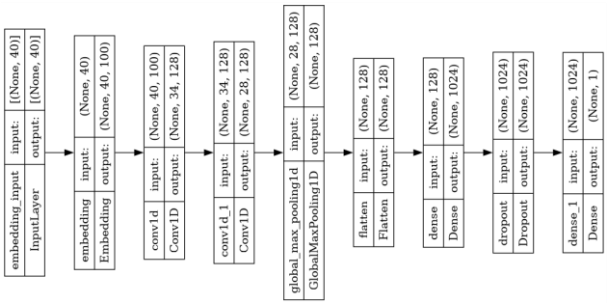
First, We chose the important features that we’ll be working on, in this case we only need the tweet and the sentiment deduced from the tweet. Second, The sentiment column is composed of Positive , negative , neutral thus we need to turn these results in to categorical in order to be able to work on them using deep learning techniques, when looking at the result of categorizing the sentiment column , it is composed of 0 ,4. Zero for indicating negative and 4 indicating positive. To make things easier we replace all 4’s by 1 so the result is [0,1]. We also shuffled the data to ensure randomness is introduced into our dataset.

3.1.2 NLP Preprocessing

First, we cleaned the tweets column to get rid of unnecessary text like mentions, URLs and special characters using Regular expressions. Second, we used a tokenizer to tokenize our cleaned tweets and convert each word to a token. Third, we converted these tokens to sequences of integers in order to be able to train our model on them. Fourth, We pad sequences to ensure that all input sequences have the same length. It pads or truncates the sequences to a specified maximum length.

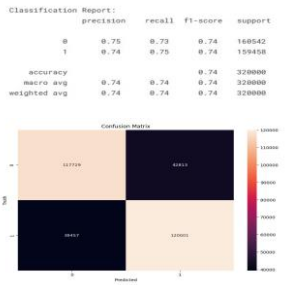
4 Proposed Model

we used CNNs in our sentiment analysis model.As CNNs can be used for text classification,text analysis,sentiment analysis and document classification as they can recognize and learn local patterns and linguistic features in text data, so they are useful in our case.We used various performance measures like Accuracy, Precision, Recall ,Confusion matrix and F1 score to evaluate our model’s performance in many ways.We used an embedding layer to map our tokens/inputs to vectors of fixed size thus saving lots of memory and making learning easier .Using Conv1D layers for text data is really convenient because Conv1D layers capture local patterns and dependencies in sequences of words.Adding a Dense layer at the end helps improve the accuracy of the model and facilitate its learning.Also using dropout layers helps the model avoid overfitting.



5 Results

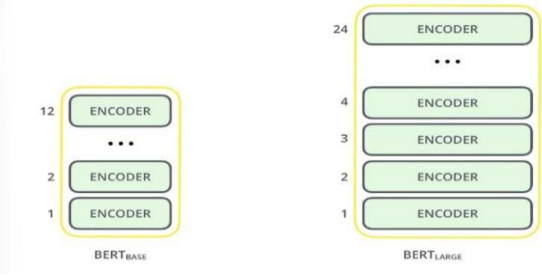
Using the architecture in the above figure gives us an accuracy of 74 percent which is good considering that the architecture is very simple and has few layers in it. The accuracy will improve by adding some complexity to the model and hyper-tuning the parameters.



6 BERT

6.1 What is BERT?

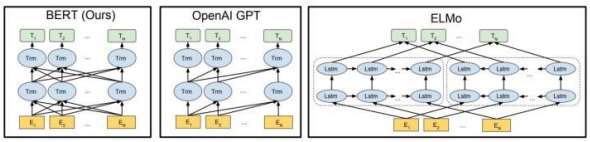
Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained model that was invented by google in 2018 and released for public use in 2019. BERT was trained on Billions of sentences to differentiate between right and wrong and it was also trained on all of Wikipedia in addition to tens of thousands of books which amounts to 3.3 billion word.BERT is essentially used in google search engine which made a huge leap in search engines field as BERT could understand what you want with few words.BERT also uses bidirectional encoder to understand the context of the sentence in which a word meaning can differ from a sentence to another.



6.2 BERT's Architecture

BERT base architecture consists if 12 layers that sums up to 110 million parameters but there is a larger model of BERT called BERT Large that consists of 24 layers and sumps up to 340 million parameter.

The above figure compares BERT's base and BERT's Large



This figure compares the embedding layer from various Langauge models like GPT ,ELMo and BERT

References

1. What is Sentiment Analysis? - Sentiment Analysis Explained - AWS (amazon.com)
2. Sentiment Analysis of Twitter Data Yili Wang 1,2,3 , Jiaxuan Guo 1,2, Chengsheng Yuan 1,2 and Baozhu Li 4,
3. Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach Yuxing Qi1 · Zahratu Shabrina2,
4. Twitter Sentiment Analysis Aliza Sarlan1 , Chayanit Nadam2 , Shuib Basri3 Computer Information Science Universiti Teknologi PETRONAS Perak, Malaysia
5. Sentiment analysis - Wikipedia
6. Sentiment Analysis: Concept, Analysis and Applications — by Shashank Gupta — Towards Data Science
7. What Is Sentiment Analysis (Opinion Mining)? — Definition from TechTarget
8. Social Media Sentiment Analysis: Tools and Tips for 2023 (hootsuite.com)
9. Sentiment Analysis — Comprehensive Beginners Guide — Thematic — Thematic (getthematic.com)
10. Sentiment Analysis — Sentiment Analysis in Natural Language Processing (analyticsvidhya.com)
11. The Importance of Social Media Sentiment Analysis — Sprout Social
12. What is Sentiment Analysis? Tools and Uses (spiceworks.com)
13. Sentiment Analysis - Lexalytics
14. Sentiment Analysis – What Is It and Why Does It Matter? (nvidia.com)
15. Sentiment Analysis: What is it and how does it work? (awario.com)
16. Sentiment140 dataset with 1.6 million tweets — Kaggle