ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology
(IJARTET) Vol. 10, Issue 8, August 2023

A review paper on Use of BERT Algorithm in Twitter Sentiment Analysis

Samruddhi Manoj Mahalle, Dr.V.S. Gulhane, Dr. Avinash.P.Jadhao
M.E. Second Year, Computer Science & Engineering, DRGIT&R, Amravati, India
Professor & Head, Information Technology, Sipna's COET, Amravati, India
Associate Professor, Department of Computer Science & Engineering, DRGIT&R, Amravati, India

Abstract: The tweet sentiment analysis is still focused on conventional messages, such as film reviews and product reviews, while significant improvement has been made as deep learning becomes wide spread, and comprehensive data sets are accessible for training (far from just emoticons and hash tags). Nevertheless, prior opinion analysis experiments typically performed on tweets, i.e., only two forms of global polarities (i.e. optimistic and negative) occur with their work/validation/test data sets. What is more, systems' judgments are not actively aligned with the specified appraisal objects. In the proposed system, we have discussed some Yolo deep learning approach for twitter sentiment analysis. We also trained our model using Yolo to get some good accuracy results.

Keywords: BERT, Sentiment, Twitter, Django, Deep Learning

I. Introduction

There has been a massive increase in the usage of micro blogging sites like Twitter in the past few years. Businesses and media organizations are increasingly trying to figure out ways to explore Twitter for knowledge about what individuals think and feel about their goods and services, driven by that development.

Companies such as **Twitratr, tweetfeel, and Social Note** are only a few that as one of their offerings promote Twitter sentiment analysis. There exist a decent content of research for identifying the various sentiments that are communicated across various domains like online reviews and news stories. Yet, a few contents of research has been conducted for identifying the various sentiments are communicated considering the informal speech and message-length restrictions of online media. Apps like automated element tags and techniques like sentimental lexicons have proven effective in many other fields for sentiment analysis, and yet can they still prove beneficial on Twitter for sentiment analysis?

The enormous range of subject matter discussed is yet another micro blogging issue. Saying that users tweet on everything and anything is not a misconception. Therefore one needs a framework for easily defining details that could be utilized for learning is required to be allowed to develop a framework for mining Twitter's feelings on any given subject. Thus, this study discusses one strategy for constructing such information utilizing hash tags that are utilized in Twitter to classify various sentiments like positive, negative, and neutral and utilize various-way emotion classifiers for learning.

Sentiment analysis is the automated mechanism for interpreting a written or spoken thought about a certain topic. Sentiment analysis has been a crucial method for interpreting this data in an environment where daily we produce 2, 5-quintillion bytes of data. It helps businesses to obtain crucial knowledge and optimize operations of all sorts [1]. Sentiment Analysis is often referred to as is Opinion Mining. This builds structures to classify and collect views inside documents. The research is a field of natural language processing (NLP) [1].

The rest of article is organized as follows. Section II describes motivation of the study. Section III Methodology. Section IV Flow chart of the implementation.

II. Motivation

Twitter sentiment analysis has lots of benefits and wide application area. Not restrictively but selectively, its exemplary applications and benefit would be described. Organization could design a successful marketing strategy with more data and insights learned via sentiment analysis. Customer's **negative or positive** messages can be used to gauge the success of the strategy.

Analysis of various posts in social networking site-based sentiments are critical for tracking abrupt changes in consumer feelings, predicting when

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 10, Issue 8, August 2023

concerns are on the increase, and taking steps before issues worsen. In real-time, one can track brand references on social networking site-based user posts with sentiment analysis and obtain valuable apprehensions.

The particular brands related to a large number of posts on networking sites can be analyzed automatically instead of manual analysis. As networking site information increasing and acquires useful insights by easy scaling of sentiment analysis tools.

Stop discrepancies resulting from human mistakes. From each piece of information, customer representatives may not always settle about which tag to utilize, so one can finish up with incorrect results. Rather than utilizing one system of regulations, machine learning techniques conduct sentiment analysis, so one can guarantee that all their information is reliably tagged.

III. METHODOLOGY

The sentiment analysis is a classification job. Positive, negative, neutral are three parts in the sentiment analysis. The analysis would be made via using machine learning (ML) and Natural language processing (NLP) methods,

In narrow scope of my study, use ML and IR methods for the analysis.

The methodology consists of two main steps in general. First is Dataset and Second step is Model utilized for sentiment analysis.

In this study, **dataset** considered is a combination of the information extracted from Twitter and the existing data based on movie reviews. For the extraction of information from Twitter, various steps need to be followed.

Step 1: Set up the Twitter application either in a web browser or on a Smartphone and generate an account. Usually, a web browser is highly recommended.

Step 2: Get the developer access to the Twitter application by creating an application on Twitter. Once the user gets the developer access, then Twitter provides a consumer key, consumer secret key, access token key, and access token secret key which are crucial and private for a user for extraction of the information from Twitter. **Step 3:** By importing the library tweepy we can use various functionalities to attain the authentication.

Step 4: Search the tweets using various search words that are related to movies and the date as well to obtain the data from that specific date to the date and time of the data extraction.

Step 5: The data obtained considered to be raw data combined with the existing data along with ratings in a structured manner as per the existing data. Step 6: The complete data need to be processed using the nltk library by removing the common words, punctuations, and stop words.

Step 7: Thus, the processed data considered for further analysis to classify the sentiments that are hidden in the reviews. The discussed data gathering is the major step in sentiment analysis.

Once data gathered and processed then the remaining part is easier to analyze.

Now the main second step discuss below that is Model utilized for sentiment analysis.

The neural network model is a pre-trained based on natural language processing utilized for sentiment analysis and that model is popularly known as BERT. The full form of BERT is "Bidirectional Encoder Representation from Transformers". In simple terms, it is a combination of a neural network as well as natural language processing. Natural language processing can also be considered as NLP corresponds to a branch of AI that interacts with linguistics to enable computers to learn and understand the communications that humans naturally considered. The major applications of NLP are social listening, sentiment analysis, and word suggestions, chatbot, and so on. The way of training the models conventionally got broken by the BERT model.

The training in this particular model will be done in a to and forth way (from first to last and again from last to first). A complete sequence of words will be passed onto the model and the model will get trained as mentioned previously. Instead of only the word that primarily precedes or follows it, BERT enables the language framework to understand word meaning based on contextual terms.

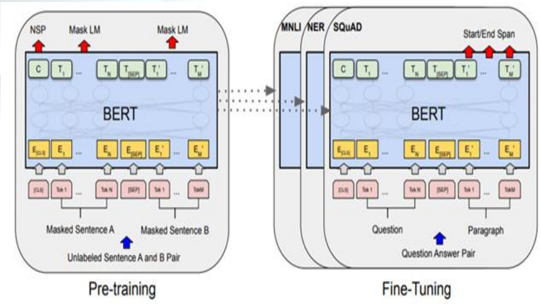

Fig. 1: Overall pre-training and fine-tuning procedures for BERT

There are two steps in our framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 10, Issue 8, August 2023

tasks. For fine tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. The question-answering example in Figure 1 will serve as a running example for this section.

A distinctive feature of BERT is its unified architecture across different tasks. There is minimal difference between the pre-trained architecture and the final downstream architecture. BERT's model architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in Vaswani et al. (2017) and released in the tensor2tensor library.1

Because the use of Transformers has become common and our implementation is almost identical to the original, we will omit an exhaustive background description of the model architecture and refer readers to Vaswani et al. (2017) as well as excellent guides such as "The Annotated Transformer." A visualization of this construction can be seen in Figure 2. BERTBASE was chosen to have the same model size as OpenAI GPT for comparison purposes. Critically, however, the BERT Transformer uses bidirectional self-attention, while the GPT Transformer uses constrained self-attention where every token can only attend to context to its left. To make BERT handle a variety of down-stream tasks, our input representation is able to unambiguously represent both a single sentence and a pair of sentences in one token sequence. Throughout this work, a "sentence" can be an arbitrary span of contiguous text, rather than an actual linguistic sentence.

A "sequence" refers to the input token sequence to BERT, which may be a single sentence or two sentences packed together. We use Word Piece embeddings (Wu et al., 2016) with a **30,000 token vocabulary**. The first token of every sequence is always a special classification token ([CLS]). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. Sentence pairs are packed together into a single sequence. We differentiate the sentences in two ways. First, we separate them with a special token ([SEP]). Second, we add a learned embedding to every token indicating whether it belongs to sentence A or sentence B. As shown in Figure 1, we denote input embedding as E, the final hidden vector of the special [CLS] token as C ∈ R H, and the final hidden vector for the ith input token as Ti∈ R H. For a given token, its input representation is constructed by summing the corresponding token, segment, and position embeddings. [3] proposed a system, this system has concentrated on finding a fast and interactive segmentation method for liver and tumor segmentation. In the pre-processing stage, Mean shift filter is applied to CT image process and statistical thresholding method is applied for reducing processing area with improving detections rate. In the Second stage, the liver region has been segmented using the algorithm of the proposed method.[6] proposed a system, in which a predicate is defined for measuring the evidence for a boundary between two regions using Geodesic Graph-based representation of the image. The algorithm is applied to image segmentation using two different kinds of local neighborhoods in constructing the graph. Liver and hepatic tumor segmentation can be automatically processed by the Geodesic graph-cut based method.

IV. Proposed Methodology

How Proposed Methodology Works:

Data Collection:

The first step in any sentiment analysis project is to gather relevant data. In this case, a dataset of Twitter tweets is required, where each tweet is labeled with its corresponding sentiment (positive, negative, or neutral). There are publicly available datasets specifically designed for sentiment analysis tasks, or you can create a custom dataset by manually annotating tweets with sentiment labels.

Data Preprocessing:

Twitter data often contains noise in the form of hashtags, mentions, URLs, and special characters. Preprocessing is crucial to remove these elements and ensure that the data is cleaned and ready for analysis. Additionally, tokenization is applied to split the tweets into individual words or sub words for further processing.

Fine-tuning BERT Model:

BERT, a transformer-based language model, has shown exceptional performance in various NLP tasks, including sentiment analysis. However, pre-trained BERT models are not directly suitable for sentiment classification on Twitter data. Therefore, the BERT model needs to be fine-tuned using the labeled Twitter dataset.

Tokenization:

Tokenization is a critical step to convert the preprocessed text into tokens that the BERT model

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology
(IJARTET) Vol. 10, Issue 8, August 2023

can understand. Each token is associated with a unique numerical representation, and the tokens are combined to form input sequences for the model.

Building the Classification Model:

After fine-tuning the BERT model, it is combined with additional layers to create a sentiment classification model. Typically, a few dense layers are added on top of BERT to adapt it to the specific sentiment analysis task. The last layer of the model is a softmax layer that provides probabilities for each sentiment class.

Training the Model:

The model is trained using the labeled Twitter dataset. During training, the model learns to map input tweet sequences to their respective sentiment classes by minimizing a suitable loss function, such as cross-entropy loss.

Model Evaluation:

After training, the model's performance is evaluated using a separate validation dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, and AUC-ROC. The goal is to assess how well the model generalizes to new, unseen data.

Hyperparameter Tuning:

Hyperparameters, such as learning rate, batch size, and number of training epochs, play a crucial role in the performance of the model. Iterative hyperparameter tuning is performed to optimize the model's performance on the validation dataset.

Sentiment Analysis on New Tweets:

Once the model is trained and evaluated, it can be used for sentiment analysis on new, unlabeled tweets. The preprocessing steps applied during training are also applied to the new data, and the fine-tuned BERT model predicts the sentiment for each tweet.

Interpretation and Application:

Finally, the sentiment analysis results can be interpreted and applied for various purposes, such as understanding public opinion, tracking sentiment trends, brand reputation management, and real-time monitoring during events. Nevertheless, the utilization of BERT in sentiment analysis has demonstrated promising results and has become a widely adopted approach in the NLP community.

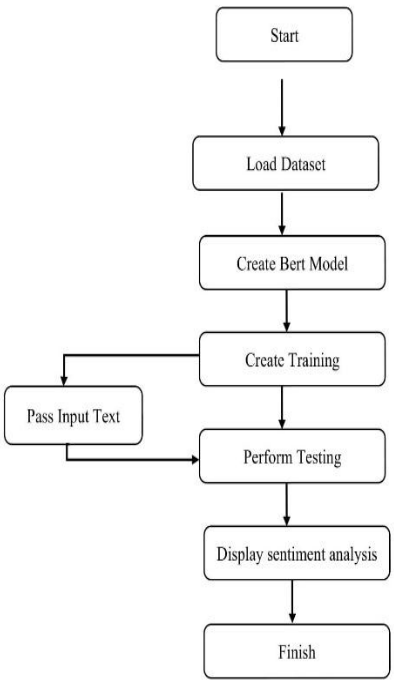

Fig 2- Flow Chart for Sentiment Analysis

V. FLOW CHART REPRESENTATION

This flow chart representation helps to see, how the project model get work with proper steps**:**

The complete implementation of the proposed framework can be implemented in various steps as mentioned in figure-3. These steps were mentioned as follows:

1. **Gathered Data:**The data is a mixture of Twitter data as well as the existing data for attaining the larger data. Due to this, the size of the data reached 25,000 instances.
2. **Data pre-processing:**Once the data gathered, it needs to be pre-processed to attain uniformity in the form of data that the analyzer looking for. Particularly, natural language processing involving project will have these three steps for the pre-processing aspects as mentioned as follows:

   a. Tokenization: It is the process of dividing the meaningful words, characters, sub words, or some symbols from the existing statements.

   b**.** Stop Words Removal: The words, for an instance, is, an, a, on, in, and so on, considered as stop words. These words will not give a meaning separately but these are used to enhance the structure of the statement. So, these are useless for our analysis. Hence, we will remove these stop words.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 10, Issue 8, August 2023

c. Stemming and Lemming**:** Stemming and lemming are almost one another represents similar working but they differ minutely. Stemming as well as lemming is getting the basic word from the existing word, but stemming won't consider the meaning of the word obtained whereas lemming will consider the meaning of the word obtained. So it is always essential that lemming should be implemented along with stemming.

3. Extraction of features: The processed data need to pass into the proposed model to get identify the features and their mapping for the identification of various sentiments depending on the dataset.

4. Identifying the keywords and classifying the movie ratings according to the recognized sentiments.

5. The training dataset and testing dataset need to pass into the proposed model and evaluate the model.

6. Now, the validation dataset will be passed on to the trained and tested model to attain the prediction or identification of polarized sentiments as positive or negative.

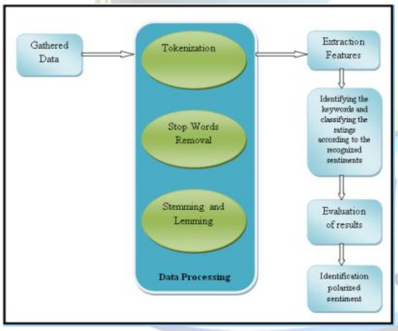

Fig 3: Flowchart representing the proposed framework

Various Stages of Twitter sentiment analysis:

The twitter-based sentiment analysis consists of various crucial steps to attain the sentiments that hidden the text format. Those steps can be represented as follows:

1. Gathering the information from the social networking site, Twitter using keywords as well as the tags.

2. Process the obtained data into a structural format so that it would be ready for further analysis.

3. Create a model that withdraws the hidden sentiments in the information obtained from Twitter.

4. Visualization of the results based on the classification of sentiment aspects is necessary for the obtained results.

REFERENCES

1. Alsaeedi, Abdullah, and Mohammad Zubair Khan. "A study on sentiment analysis techniques of Twitter data." International Journal of Advanced Computer Science and Applications 10.2 (2019): 361-374.
2. Chen, Xin, et al. "A novel feature extraction methodology for sentiment analysis of product reviews." Neural Computing and Applications 31.10 (2019): 6625-6642.
3. Christo Ananth, D.L.Roshni Bai , K.Renuka, C.Savithra, A.Vidhya, "Interactive Automatic Hepatic Tumor CT Image Segmentation", International Journal of Emerging Research in Management &Technology (IJERMT), Volume-3, Issue-1, January 2014,pp 16-20.
4. Haque, TanjimUl, Nudrat Nawal Saber, and Faisal Muhammad Shah. "Sentiment analysis on large scale Amazon product reviews." 2018 IEEE International Conference on Innovative Research and Development (ICIRD). IEEE, 2018.
5. Ireland, Robert, and Ang Liu. "Application of data analytics for product design: Sentiment analysis of online product reviews." CIRP Journal of Manufacturing Science and Technology 23 (2018): 128-144.
6. Christo Ananth, D.L.Roshni Bai, K.Renuka, A.Vidhya, C.Savithra, "Liver and Hepatic Tumor Segmentation in 3D CT Images", International Journal of Advanced Research in Computer Engineering & Technology (IJARTET), Volume 3,Issue-2, February 2014,pp 496-503.
7. Kiran, M. Vamsee Krishna, et al. "User-specific product recommendation and rating system by performing sentiment analysis on product reviews." 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2017. [8] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." Fifth International AAAI Conference on weblogs and social media. 2011.
8. Liang, Ruxia, and Jian-Qiang Wang. "A linguistic intuitionistic cloud decision support model with sentiment analysis for product selection in E-commerce." International Journal of Fuzzy Systems 21.3 (2019): 963-977.
9. M. M. Mostafa, "Clustering halal food consumers: A Twitter sentiment analysis," Int. J. Mark. Res., vol. 61, no. 3, pp. 320–337, 2019.
10. Mubarok, Mohamad Syahrul, Adiwijaya, and Muhammad DwiAldhi. "Aspect-based sentiment analysis to review products using Naïve Bayes." AIP Conference Proceedings. Vol. 1867. No. 1. AIP Publishing LLC, 2017.
11. Ramadhani, A. M. and H. S. Goo, "Twitter sentiment analysis using deep learning methods," Proc. - 2017 7th Int. Annu. Eng. Semin. Ina. 2017, 2017.

