# Twitter Sentiment Analysis Using TensorFlow

ABSTRACT

Twitter is a social networking platform used for expressing their views about politics, products, sports etc. Hence, tweets can be used as a valuable source for mining public's opinion.

Today due to rapid increase in the data on social media platforms like Twitter, that it is hard to keep a track on customer reviews and opinions about their products and services. Sentiment analysis is a process of identifying whether a text expresses positive, negative opinion. The NLP (natural language processing) algorithms are used for sentiment analysis in the proposed work. The sentimental analyzer is used for overcoming the challenges to identify the twitter tweets text sentiments (positive, negative) by implementing neural network using tensorflow.

The objective of this paper is to give step-by-step detail about the process of sentiment analysis on twitter data using neural network. This paper also provides details of proposed approach for sentiment analysis.

## Introduction

Millions of people are using Twitter and expressing their emotions like happiness, sadness, angry, etc. The Sentiment analysis is also about detecting the emotions, opinion, assessment, attitudes, and took this into consideration as a way humans think. Sentiment analysis classifies the emotions into classes such as positive or negative. Nowadays, industries are interested to use textual data for semantic analysis to extract the view of people about their products and services. Sentiment analysis is very important for them to know the customer satisfaction level and they can improve their services accordingly. To work on the text data, they try to extract the data from social media platforms. There are a lot of social media sites like Google Plus, Facebook, and Twitter that allow expressing opinions, views, and emotions about certain topics and events. Microblogging site Twitter is expanding rapidly among all other online social media networking sites with about 200 million users. Twitter was founded in 2006 and currently, it is the most famous microblogging platform. In 2017 2 million users shared 8.3 million tweets in one hour. Twitter users use to post their thoughts, emotions, and messages on their profiles, called tweets. Words limit of a single tweet has 140 characters. Twitter sentiment analysis based on the NLP (natural language processing) field. For tweets text, we use NLP techniques like tokenizing the words, removing the stop words like I, me, my, our, your, is, was, etc. Natural language processing also plays a part to preprocess the data like cleaning the text and removing the special characters and punctuation marks.Sentimental analysis is very important because we can know the trends of people's emotions on specific topics with their tweets.

We are using python language in the implementations and Jupyter Notebook that support the machine learning and data science projects. Here we create a model of sentimental analyzer for overcoming the challenges to identify the twitter tweets text sentiments (positive, negative) by implementing neural network using tensorflow.

## Data Description

The dataset considered for the analysis is the famous sentiment140 dataset. It is a public Dataset that contains 1,600,000 Tweets extracted using the twitter API . Tweets are classified as positive or negative.

**URL:**
**[https://github.com/h2oai/social_ml/blob/master/sentiment_m140_.csv](https://github.com/h2oai/social_ml/blob/master/sentiment_m140_.csv)**

The data given is in the form of comma-separated values files with tweets and their corresponding sentiments. The training dataset is a csv file of type tweet_id, sentiment,tweet where the tweet_id unique and emoticons contribute to predicting the sentiment, but URLs and references can be ignored. The words are also a mixture of misspelled words, extra punctuations, and words with many repeated letters. The tweets, therefore, have to be pre-processed to standardize the dataset.

## Methodology and Implementation

### 1.Text Preprocessing:

Twitter data may be in unstructured format that is not good for extracting feature. Tweets may consist of empty spaces, stop words, slangs, special characters, hashtag, emoticons, time stamps, abbreviations, URL's etc. For mining these data we should have to pre-process the data first by the using the functions of NLTK. While doing pre-processing our first aim is to extract message then we will remove all hashtags(#), empty spaces, repeating words, stop words(such as he, she, them, the etc.). Emoticons and abbreviation will be replaced by their corresponding meaning such as :-), =D, LOL. They will be replaced by happy, laugh and laughing out loudly respectively. After done this thing we are ready to give this pre-process data to our new classifier for further process so we could get our required result. We did code in python where we define function which would be used to get processed data.

*Lowercasing:* Convert all text to lowercase to ensure consistent analysis.

*Removing Stop Words:* Eliminate common word (e.g., and, the) that don't carry much sentiment information.

*Removing punctuations:* We will clean and remove the punctuations because these are the noise in the data and not meaningfull.

*Removal of emoticons:* Replace emoticons with their correct meaning.

*Removal of duplicates:* We will clean and remove repeating characters in the words

*Remove emails*: To remove the @ symbol, delete @ together with the username.

*Remove URL's:* URL stands for uniform resource locator. offers options to remove URLs .

*Remove numbers:* We will clean and remove the numbers in the data

*Tokenization:* Break the text into individual words or tokens.

*Stemming and Lemmatization:* Reduce words to their root form to normalize variations (e.g., "running" to "run").

### 2.Selecting a Model:

Here we Implementing Tensorflow based Neural Network model for training. The text data is tokenized using the Tokenizer class from Keras, which converts words into numerical sequences. A maximum sequence length of 500 words is set, and sequences are padded or truncated to this length using sequence.pad_sequences().The data is split into training and testing sets using train_test_split( ) from sklearn, with a test size of 30%. The model architecture is implemented using TensorFlow and Keras. The input layer takes sequences of length 500. The LSTM (Long Short-Term Memory) layer processes the sequential data, capturing dependencies between words in the tweets. The model is compiled using binary_crossentropy as the loss function since it's a binary classification problem. A batch size of 80 and 6 epochs are chosen for training. Overall, this model leverages word embeddings and LSTM layers to capture the semantic meaning and sequential dependencies in tweets, followed by dense layers for classification. Dropout is used to prevent overfitting, and the model is trained using a specified batch size and number of epochs.

### 3.Model Evaluation:

Assess the model's performance using metrics like accuracy on a separate test dataset. Accuracy is the number of correctly classify tweets from all the tweets of positive and negative. Apply the trained model to new, unseen text to predict the sentiment. As the model give probabilties so we are setting a threshold 0.5. More than 0.5 will be the positive tweets and lower will be negative tweets.

## Result

The model achieves an accuracy of approximately 75% on the test data, indicating its ability to correctly classify sentiments.The confusion matrix reveals the distribution of correct and incorrect predictions, helping to identify areas of improvement.The ROC curve provides a visual representation of the model's performance across various thresholds, indicating its discriminative ability.Overall, the implementation demonstrates the use of LSTM networks for sentiment analysis and provides insights into the model's performance through evaluation metric.

## Conclusion

In today’s world, spacious amount of data is generated by various communication such as social media, organizations etc. these data may or may not be in structured form. Therefore to understand the polarity of data first we need to do the sentiment analysis of data. Opinion mining can be performed in various field such as marketing and customer feedback. Large number of organizations are taking the valuable feedback of person and performing opinion mining on those data so that they could provide the better services to the customer and this data helps the organizations to enhance their future services. Furthermore, there are various scopes where we can perform the opinion mining such as sentence, paragraph, documents, sub sentences levels.

Basically our goal is to find the public opinion and perform the opinion mining. Generally what happens, through tweets people express their thoughts, feelings etc. but we could not able to find the people thoughts and feelings. So by performing sentiment analysis on those tweets finally we can conclude how many person are in favor of this mission and how many person are against of this mission. Future scope includes, Developing a web application for opinion mining can make the process more accessible and user-friendly. Improving classifier systems to handle sentences with multiple meanings can enhance the accuracy of sentiment analysis. Expanding classification categories can provide more nuanced results. Integrating image processing techniques to detect images in tweets can enrich the analysis by considering visual content along with text.

## References

Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009), p.12*

Mukherjee S., Malu A., Balamurali A.R, Bhattacharyya P.“TwiSent: A Multistage System for Analyzing Sentiment in Twitter”.

Neethu, M., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). (2013).