Sentiment Analysis of Twitter Data

Om Trivedi[1], Harmeet Kaur[2]

[1]B.Sc. Physical Science (Computer Science) II year, Hansraj College, University of Delhi 110007, Delhi, India, umeshtrivedi339@gmail.com

[2]Department of Computer Science, Hansraj College, University of Delhi 110007, Delhi, India, hkaur@hrc.du.ac.in

Abstract

The Internet has become a platform for online learning, exchanging ideas and sharing opinions. The fast-growing social web contributes substantial amount of data generated by users in the form of reviews, comments, tweets, posts and opinions. People are using social networking sites to voice their opinions regarding daily issues. Even though the content holds a lot of potential but mining such humongous amount of data requires a lot of time and effort hence, there is a need to develop an intelligent system which in no time can analyze the content and classify them into categories being positive, negative or neutral. Twitter is one of the social media that is gaining popularity, it is a perfunctory platform that generates a constant amount of data of events having major or minor concern, around the world. Twitter offers organizations a fast and effective way to analyze customers' perspectives toward the critical to success in the marketplace. An analysis of Twitter may, therefore, give insights of the sentiments generated in the population regarding various events.

Opinion investigation of Twitter data is a field that has been given much attention over the last decade and involves dissecting “tweets” and the content of these expressions. In this paper we focus mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. A comparative analyses of different Natural Language processing (NLP) algorithms for sentiment analysis are presented in this paper. The objective of this research paper is to contribute to an understanding of the actual and potential role of sentiment analysis techniques of Textblob and RoBERTa by applying them on the Twitter data. The paper is organized as follows: the first two sections comment on the definitions, motivations, and classification techniques used in sentiment analysis. Finally, discussions and comparisons of the latter are highlighted. Furthermore, we also discuss directions for future research on how twitter sentiment analysis can utilize theories and technologies for other fields such as cognitive science, semantic web, big data and visualization.

Keywords: Twitter, Data Analysis, Sentiment analysis, Machine learning, Natural Language Processing.

Introduction

Sentiment Analysis is often referred to as subjective analysis, opinion mining, and appraisal extraction. Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It is also known as “opinion mining” and alludes to the utilization of natural language processing (NLP), text mining and computational linguistics to methodically recognize, extricate, evaluate, and examine emotional states and subjective information. It is the automated mining of attitudes, opinions, emotions from text, speech and database sources. It involves classifying opinions into positive, negative or neutral. Sentiment analysis can be done on a document level, sentence level, aspect level or feature level. There are two major approaches to sentiment analysis: machine learning based, which uses classification to classify text into various categories, and lexicon based, which uses meaning of the dictionary to classify words as positive, negative or neutral and to what extent the values are positive or negative. The aim of this paper is to see how sentiment analysis of tweets can help determine the general sentiment of the masses over a certain happening. This can give some insights into the future like what changes need to be made or the general likes or dislikes of the target audience. The objective of this research paper is to contribute to an understanding of the actual and potential role of sentiment analysis techniques and analyze its results by its application on the Twitter data.

Sentiment in general means feelings that are emotions, attitudes and opinions. These are subjective matters and not facts. Unlike facts, sentiments are ever changing and can be molded in one's favor if required. The sentiments are influenced by a variety of factors like thinking of other beings, ideas and opinions in the surroundings. In ancient times, the media, rumors and statements of the public moved the opinions of the masses. In this tech-oriented age where everyone and everything is on the web, public opinion is greatly influenced by what is buzzing and trending on the internet.

Methodology

Sentiment Analysis involves following steps which involves:

1. Data Collection
2. Data Cleaning
3. Classification
4. Model Evaluation

1. Data Collection

In this project, Twitter Data is used for sentiment analysis. To connect to Twitter and query the latest tweets, first create an account on twitter and an application which can interact with Twitter API. For this go to apps.twitter.com/app/new and generate the API keys.

Twitter provides REST APIs, which can be used to interact with their service; Tweepy is one such API which is used here. It is an open-source Python package that gives a very convenient way to access the Twitter

API with Python. Tweepy includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details, such as: data encoding and decoding.

For this project 100 tweets from Mr. Elon Musk's Twitter account were used for experimentation.

2. Data Cleaning

Data Cleaning means preprocessing of data. It involves removing stop words and special characters like @, '#', blank spaces etc.

These could be summoned up as follows:

- Hash tags: they are very common in tweets. Hash tags represent a topic of interest about which the tweet is being written. Hashtags look something like #topics
- @Usernames: they represent the user mentions in a tweet. Many times, a tweet is written and then is associated with some twitter user, for this purpose these are used.
- Retweets (RT): retweets are used when a tweet is posted twice by the same or different user.
- Emoticons: these are very commonly found in tweets. Using punctuations facial expressions are formed to represent a smile or other expressions, these are known as emoticons.
- Stop words: stop words are those words which are useless when it comes to sentiment analysis.
- Stemming is also performed as a part of preprocessing technique.

2.1 Tokenization

It is one of the most important steps in text analysis. Tokenization is used in natural language processing to split paragraphs and sentences into smaller units that can be more easily assigned meaning. The first step of the Natural Language processing (NLP) process is gathering the data (a sentence) and **breaking** it into smaller units called tokens which are understandable.

3. Classification

There are mainly two classifications techniques for sentiment analysis for the twitter data:

1. Machine Learning Approach
2. Lexical based Approach

For this work, a Lexical based approach has been used.

3.1 Lexical based approach

One of the approaches or techniques of semantic analysis is the lexicon-based approach. This technique calculates the sentiment orientations of the whole document or set of sentences(s) from semantic orientation of lexicons. Semantic orientation can be positive, negative, or neutral.

The dictionary of lexicons can be created manually as well as automatically generated. The WorldNet dictionary is used by many researchers. First, lexicons are found from the whole document and then WorldNet or any other kind of online thesaurus can be

used to discover the synonyms and antonyms to expand that dictionary.

Lexicon-based techniques use adjectives and adverbs to discover the semantic orientation of the text. For calculating any text orientation, adjective and adverb combinations are extracted with their sentiment orientation value. These can then be converted to a single score for the whole value.

First TextBlob was used, which is a Lexicon-based sentiment analyzer. TextBlob can be used for complex analysis and working with textual data.

When a sentence is passed into TextBlob it gives two outputs, which are polarity and subjectivity. Polarity is the output that lies between [-1, 1], where -1 refers to negative sentiment and +1 refers to positive sentiment, 0 refers to the neutral sentiment.

Exploratory analysis was performed on tweets of Mr. Elon Musk. Figure 1 shows these results in a wordcloud form which shows that that two of the most commonly used words by him are Twitter and work. Figure 2 shows that most of the tweets are above polarity 0.0 i.e., more than 50% tweets are neutral.

The pie chart of the tweets in Figure 3 shows sentiment of tweets predicted by TextBlob:

50.5% of the Tweets are Neutral
7% of the Tweets are Negative
42.5% of the Tweets are Positive

For this project, RoBERTa-based model has also been used which is trained on ~58M tweets and finetuned for sentiment analysis with the TweetEval benchmark.

Analysis was also performed using RoBERTa (A Robustly Optimized BERT Pretraining Approach) by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. It is based on Google's BERT model which was released in 2018.

This model is built on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

FIGURE LEGENDS

Figure 1: Shows results in a WordCloud form

Figure 2: Shows the polarity of the tweets of Mr. Musk

Figure 3: Pie Chart of the tweet's sentiment predicted by TextBlob

Figure 4: Shows a Pie chart of the responses given by people

FIGURES

Figure 1: Shows results in a Wordcloud form

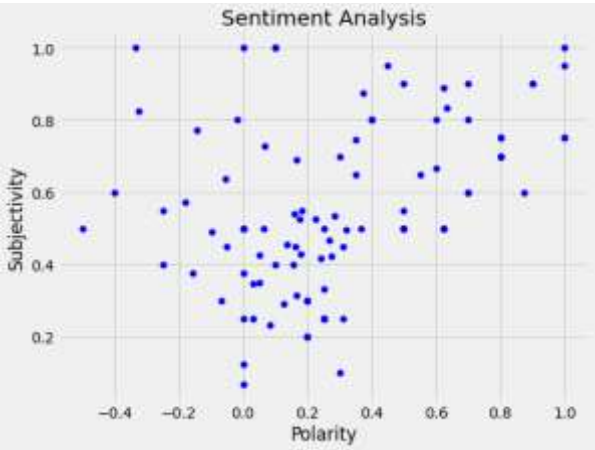

Figure 2: Shows the polarity of the tweets of Mr. Musk

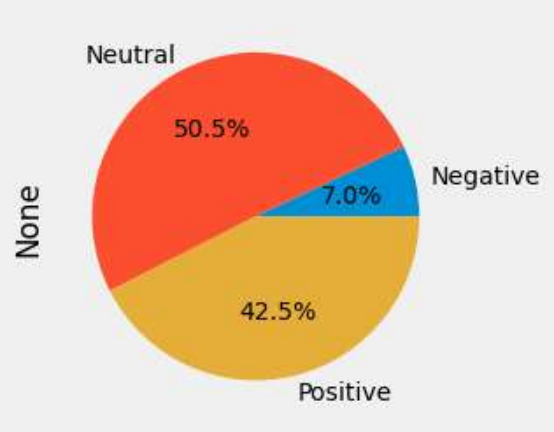

Figure 3: Pie Chart of the tweet's sentiment predicted by TextBlob

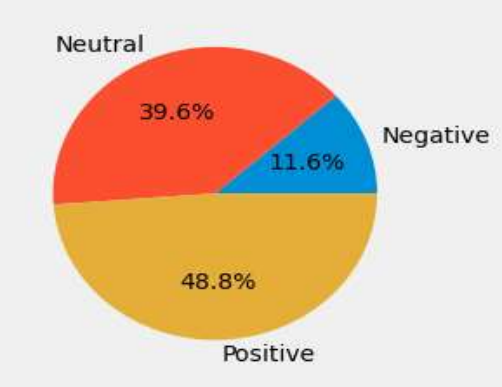

Figure 4: Shows a Pie chart of the responses given by people

Results and Observations

For comparing the accuracy of TextBlob model and RoBERTa model, a form was floated in which people had to fill whether they find the given tweet positive, negative or neutral. These responses were compared with the results reported by TextBlob model and RoBERTa model. Randomly 8 tweets were selected out of the total 100 in the dataset and collected 200+ responses from over 28 people.

Following is the summary of the responses obtained from people as shown in Figure 4:

Positive Tweets: 48.8%
Negative Tweets: 11.6%
Neutral Tweets: 39.6%

TABLES

Table 1: Results predicted by TextBlob model and RoBERTa model

Table 1 shows the results of some of the tweets by Mr. Elon Musk, along with some statistics and polarity of the tweets as per human user, TextBlob and RoBERTa models for sentiment analysis.

Tweets	Fun level on Twitter has definitely increased! Iâ€™m having a great time tbh	Good for you !	This is necessary to restore public trust	Thanksgiving cuisine is such a delightful symphony of flavor	Oh we have quite the adventure ahead!	I wonder what Earth will be like 88 million years from now	World-class software aces are joining Twitter	This makes sense
Survey top response	Positive	Positive	Positive	Positive	Neutral	Positive	Positive	Neutral
Total no. of responses	24	24	24	24	24	24	24	24
Frequency of top response	11	15	12	12	14	9	13	11
TextBlob (Model 1) Response	Positive	Positive	Neutral	Neutral	Positive	Neutral	Neutral	Neutral
RoBERTa (Model 2) Response	Positive	Positive	Neutral	Positive	Positive	Neutral	Positive	Positive
Prediction (Model 1)	Right	Right	Wrong	Wrong	Wrong	Wrong	Wrong	Right
Prediction (Model 2)	Right	Right	Wrong	Right	Wrong	Wrong	Right	Wrong

Table 1: Compares results predicted by TextBlob model and RoBERTa model with survey response of people

Conclusions

The goal of this research is to introduce the basic concepts and techniques for sentiment analysis of tweets and analyze accuracy of different models. We focused on Twitter and have implemented a Python program for sentimental analysis. In recent years,

researchers have become increasingly interested in analyzing tweets based on the sentiments they represent. This interest comes from the fact that a great number of tweets are posted on Twitter, which provides vital information on the sentiments of the public on a variety of subjects.

The greatest difficulties that were encountered were in determining the best approach for detecting sentiments in Twitter data because comparing various approaches is a highly challenging task when there is a lack of agreed benchmarks. Interesting areas for future study include the fluctuations in the performance of sentiment analysis algorithms in cases where multiple features are considered.

In future the aim is to perform similar analysis to increase the accuracy of the model and do so without extracting any features, by using different deep learning techniques.

ACKNOWLEDGEMENTS

I would like to heartily thank Hansraj College for providing the platform for this research. I would also like to acknowledge all the secondary sources and friends who provided us the information in the course of this research.

References

1. Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. Computer Speech & Language, 44:61–83.
2. R. Parikh and M. Movassate, “Sentiment Analysis of User-GeneratedTwitter Updates using Various Classi_cation Techniques", CS224N Final Report, 2009
3. Go, R. Bhayani, L.Huang. “Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper, 2009
4. Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010
5. Neethu M.S. and Rajashree R, “Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013, at Tiruchengode, India. IEEE – 31661
6. TextBlob, 2017, https://textblob.readthedocs.io/en/dev/
7. Dasgupta, S. S., Natarajan, S., Kaipa, K. K., Bhattacharjee, S. K., & Viswanathan, A. (2015, October). Sentiment analysis of Facebook data using Hadoop based open-source technologies. In Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on (pp. 1-3). IEEE
8. Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, 31, 527-541.