A Exploration of Sentiment Analysis on Twitter Data Sets

Shilpi Kanojia [1] , Bhagya Laxmi [2]
M. Tech. (CSE) [1] , Asst. Prof. Dept. of Computer Science [2]
GRD Institute Of Management & Technology, Dehradun [1,2]

Abstract-Currently, social networks play an important role in sharing data and sharing their ideas. The emotional effects of a person play an important role in their daily life. The analysis of feeling is the process of analyzing the ideas and polarity of people. Twitter is the main platform to share ideas, opinions and emotions on different occasions. The sentiment analysis of Twitter is a method to analyze the emotions of the tweets (messages posted by users in the tweets). Tweets help extract user sentiment values. The data provides an indication of polarity, such as positive, negative or unbiased values. It focuses on human tweets and hashtags to understand all aspects of the standard**.**

Keywords- Sentiment Analysis, Twitter, Tweets, python

1. Introduction

When studying seriously and make daily decisions, it often sees people's thoughts. During the political election, he advised the forum of political debate, reading customer reports during the purchase of customers and asked friends and suggested dinner for dinner. Today, the Internet can find millions of people from the latest gadgets to a political philosophy. According to the latest Internet and Civic Joining Survey, "One of five people is posting content on social networking sites for political or social issues, or some citizenship or political participation.

1. In another study, one third (33%) blogs of internet users were reading blogs, 11 of which were doing every day.

2. The Internet was discussing rapidly and to provide information to become a forum for people.

To prepare, create a new area for text analysis and enhances the research topics related to the traditional fact and to convey knowledge-related applications with the ideas of text emotions. In the past decade, a wide range of attention has been received in the industry and academy to extract text emotions. Most companies are recognizing the importance of awareness about product and services to Internet users. This article covers the field of article analysis, which includes article articles. The top priority is to define emotions and define their relationship with the text.

1.1 Sentiment Analysis through NPL

The analysis is intended to extract a natural language processing and information that aims to analyze a large number of positive or negative feedback, questions and documents according to the author. Usually, the objective of emotional analysis is to determine the behavior of a speaker or author for the overall integrity of a subject or document. In recent years, rapid growth in internet usage and public opinion exchange has become a driver's power of emotional analysis today. Web is a large collection of organized and non-organized data. This data analysis is a difficult task to extract potential public opinion and feelings.

An emotional analysis based on feelings in the document can be changed for positive, negative or purpose. This phrase can be based on which the text is classified as emotions. The SA may be based on a paragraph, and the phrase is done according to the proparity. Passion analysis identifies the phrase with specific emotions in the text. The author can talk about some objective facts or mental opinion. They need to make a difference between the two. SA finds the subject that was the method of analysis. The text may contain many institutions, but it should be discovered which emotions should be directed. It sets the level of polish and emotion. Emotions can be ranked as objective (realistic), positive (part of happiness, prosperity or author's satisfaction) or negative (depressed, frustration or depression representation for the part of the author). The emotional score can be given more on the basis of encouragement, negative or neutrality.

1.2 Objective of the Research

This data is very important to understand this question before processing. The problem statement is as follows: The purpose of this task is to detect hate speeches in tweets. For simplicity, he says that there is a racism or hunger absorbed in it, so it includes a tweet of hatred. Therefore, work is to rate racist or sexual tights from other tweets. Typically, tweets and labels have been given a training model, where Label '1' Tweet is related to Nigeria/sexually labeled and Label '0' says that Tweet is not Nigerians /sexually explicit, intended for the purposes of test data Labs on the base are to be predicted.

- To collect the data in CSV file as trend data and test data sets
- To perform the analysis over the data
- To detect negative tweets (racist/sexist/bigot) from the test data sets
- To find the percentage accuracy of the tweet outputs

The objective of this task is to discover the hate speech on Twitter. Say that a geek contains a hate speech if it has a racist or partial feeling associated with it. Therefore, the task is to classify racist or sexual tweets from other tweets. Formally, if you provide a form of Twitter training and labels, where the "1" label indicates that Twitter is racist / partial and "0" indicates that Twitter is not racist / partial, your goal is to predict the labels in the data set test.

2. Methodology

Explains the methodological steps followed in this work. In the first step, we gathered data from Twitter. It then describes the process of aggregating schedules, communication between users and Twitter user profiles. A schedule analysis was conducted for each user to determine what policy tweets are and what is not. Details of this identification process are described. In the proposed methodology, we analyzed the feelings of tweets that

have a different approach: one for racist and non-racial tweets that dealt with both candidates at the same time and others with Twitter.

2.1 Tweets Preprocessing and Cleaning

This is looking for a document in this office space. This scene has been created with less likely one to find the document easily because everything is kept in its proper place. Data cleaning exercises are exactly the same. If the data is managed in a regular manner, the right information is easy to find. Text data type is an essential step because it produces raw text ready for mining that makes it easy to extract information from the text and apply the algorithm of machine learning. If they leave this step then there is a great chance that you are working with noise and contradictory data. The purpose of this stage is to clear the noise that are less relevant to finding emotions of time, regardless of particular characters, numbers, and conditions that do not overweight in terms of text. In one of the later steps, they will be taken to remove digital features from our Twitter text data. This feature space is created using all the unique words in the entire figure. So, if they offer our data well, they will succeed in achieving a better quality feature.

2.2 Removing Twitter Handles (@user)

As mentioned above, tweets contain many Twitter handles that how twitter has been recognized on Twitter. They will remove all of these Twitter handle from the data because they do not take more information. For our convenience, allow Let's first finger train and test set. It saves trouble and trials on the train twice the test and train.

2.3 Removing Punctuations, Numbers, and Special Characters

As conversations, wings, numbers and special characters do not help much. It's better to remove them from text as they remove Twitter handle. Here they will change everything except letters and weapons with spaces.

2.4 Removing Short Words

They must be careful about choosing the length of words that they want to remove. So, I have decided to reduce all words by 3 or at least. For example, the terms "ham", "oh" are of very little use. It's better to get rid of them.

2.5 Tokenization

Now they will break all the cleaned tweets in our database. Token is individual terms or words, and Tokenize is a process to distribute a string of text.

2.6 Stemming

Stemming laws are based on word removing words ("ing", "ly", "es", "etc.").For example, for example, "sports", "player", "players", "played" and "game" words "different".

3. Story Generation and Visualization from Tweets

This section find clean leggings. The search and insight of the data, no matter whether its text or any other data is necessary to get insight into it. Do not limit yourself in this tutorial in those ways, which feel free to find as much as possible data. Before starting the investigation, it is necessary to ask for questions related to the data in hand. Some possible questions are as follows:

- What are the most common words in the entire dataset?
- What are the most common words in the dataset for negative and positive tweets, respectively?
- How many hashtags are there in a tweet?
- Which trends are associated with my dataset?
- Which trends are associated with either of the sentiments. Are them compatible with the sentiments.

4. Extracting Features from Cleaned Tweets

To analyze a suggested figure, it needs to be converted into features. Depending on the application, you can use built-in techniques to create text functions - package words, TF-IDF and word embedded.

4.1 Bag-of-Words Features

This bag-of- the package is a way of representing the text through linguistic features. Consider one of the points (words of words) named CDD documents {d1, d2 ... dd} and anonymous unique tokens. N Tokens (Words) will create a list and its size bag will be given by Matrix MD X. Each row in the Matrix M do ent d (i) contains the frequency of the tech.

Let us understand this using a simple example. Suppose they have only 2 document

D1: He is a lazy boy. She is also lazy.
D2: Smith is a lazy person.
The list created would consist of all the unique tokens in the corpus C.

= [‘He’,’She’,’lazy’,’boy’,’Smith’,’person’]

4.2 Model Building: Sentiment Analysis

Now they have completed the first modeling steps necessary to get the figures in the correct form and shape. Now they will build two models of set-off off-words and prediction models based on the TF-IDF database.
They will use logistic depression to make models. It predicts the possibility of an event occurring by fitting data at a logic function. The following equation is used in Logistic Regression:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(\text{Age})$$

Here, D=2, N=6

4.3 Experiment Analysis

They have two csv files 'train.csv' and test.csv, including "Range Index: 31 962 entries, 0 to 31961 data columns (total 3 columns): ID 31962 unnecessary int64, label 3 962 non-anger. Into 64 tweet 31 962 unusual objection "and" range index: 17197 entries, 0 to 17196 data columns (total 2 columns): ID 17197 unusual int64 tweet 17197 unusual objection ".They find 780 Racist Tweets and 16417 tweets

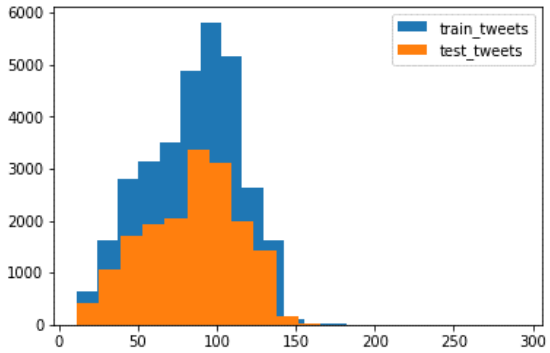

Figure 1 Train tweets vs. Test tweets

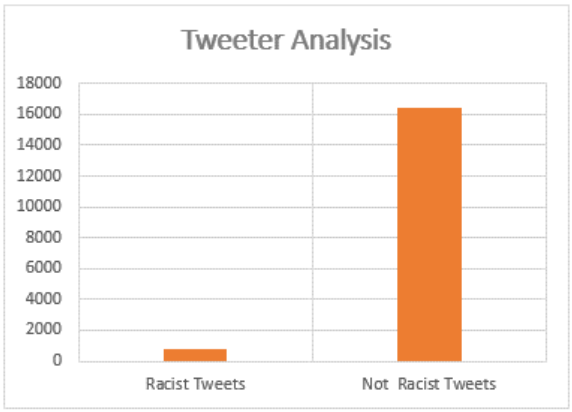

Figure 2 Compare tweets of racist vs. non racist

As the above figure shows that balanced analysis has been comb ined with wish tweets with non-wishing tweets. Find the nature o f tweets from the database given to this article. This data has bee n collected from Twitter.

5. Results and Discussion

Natural language processing (NLP) is the heat of research in these days these days, and most of the NLP applications are emotional analysis. With the opinion of making the entire marketing strategy, this domain has completely changed the business tasks, so it is an area where every data scientist must be familiar with. Thousands of hour's text documents can be processed for passions (including organizations, topics, topics, etc.) in seconds, compared to that time, to make a team manually complete the same task. Will take They will do this by following the necessary steps to address the problem of general emotion analysis. They will start with preprocessing and raw text cleaning of tweets. Then they'll discover the clear text and try to get some interference about the tweets. After that, they will remove numerical features from the data and finally use their feature set, to train the model and identify tweets.

6. Conclusion and Future Scope

This paper focus on this subject on the study of passion analysis method to find the main trend of the database. Twitter emotions are designed to analyze the customer's view for the importance of space in the market. This program is using a machine oriented learning approach that is more accurate to analyze emotions; natural language processing techniques will be used as well. Emotional analysis is a ready field with different usage applications. Although emotional analysis has to be challenged by the initial reason for the processing of their natural language, due to high demand, it has been mostly developed in the past few years. Not only companies need to know that their products and services are considered by consumers (and compared to competition), but consumers want to know others' opinions before buying decisions. Insight of product and currently in the field will help increase the growing needs of technical challenges facing emotional analysis and feedback for the final future. The next-generation feedback mining system needs a deeper bond between members of full knowledge, with human thinking and psychological approaches. It will lead to a better understanding of natural language opinion and between the extraordinary information more than this, space between human ideas and formulated data, which can be analyzed and implemented by a machine.

References

[1] Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 764-779.

[2] Singh, T., & Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, *89*, 549-554.

[3] Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

[4] Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.

[5] Cliche, M. (2017). BB_twtr at SemEval-2017 task 4: twitter sentiment analysis with CNNs and LSTMs. *arXiv preprint arXiv: 1704.06125*.

[6] Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2870-2879.

[7] Amolik, A., Jivane, N., Bhandari, M., & Venkatesan, M. (2016). Twitter sentiment analysis of movie reviews using machine learning techniques. *International Journal of Engineering and Technology*, *7*(6), 1-7.

[8] Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, *49*(2), 28.

[9] Wehrmann, J., Becker, W., Cagnini, H. E., & Barros, R. C. (2017, May). A character-based convolutional neural network for language-agnostic Twitter sentiment analysis. In *2017 International Joint Conference on Neural Networks (IJCNN)*(pp. 2384-2391). IEEE.

[10] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*, *52*(1), 5-19.

[11] Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems*, *33*(4), 1034-1058.

[12] Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access*, *6*, 23253-23260.

[13] Kanakaraj, M., & Guddeti, R. M. R. (2015, February). Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)* (pp. 169-170). IEEE.

[14] Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. S. (2015). Twitter sentiment analysis for large-scale data: an unsupervised approach. *Cognitive computation*, *7*(2), 254-262.

[15] Philander, K., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, *55*(2016), 16-24.

[16] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 451-463).

[17] Severyn, A., & Moschitti, A. (2015). Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 464-469).

[18] Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015, June). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 470-478).

[19] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)* (pp. 1-18).

[20] Giorgis, S., Rousas, A., Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2016). aueb. twitter. sentiment at SemEval-2016 Task 4: A weighted ensemble of SVMs for Twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 96-99).

[21] Rosenthal, S., Farra, N., & Nakov, P. (2017, August). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 502-518).

[22] Furini, M., & Montangero, M. (2016, June). TSentiment: On gamifying Twitter sentiment analysis. In *2016 IEEE Symposium on Computers and Communication (ISCC)* (pp. 91-96). IEEE.

[23] Agrawal, Rajagopalan, Srikant, and Xu (2003). Mining newsgroups using network arising from social behavior. *Twelfth international World Wide Web Conference*.

[24] Hu, M. and Liu, B. (2005). Mining and summarizing customer reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

[25] Hu, M. and Liu, B. (2005). Mining and summarizing customer reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

[26] Liu, B. (2006). *Web Data Mining*, chapter Opinion Mining. Springer.

[27] Melville, P., Gryc, W., and Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. *Proceedings of the Conference on Knowledge Discovery and Data Mining 2009*.

[28] Pang, B. and Lee, L. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10:79–86.

[29] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135.

[30] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135.

[31] Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

[32] R. Quirk, S. Greenbaum, G. L. and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

[33] Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

[34] Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30:277–308.

[35] Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30:277–308.

[36] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar, Thumbs up?: sentiment classification using machine learning techniques, In Proceedings of the ACL-02 conference on Empirical Methods in Natural Language, 2002