Sentiment Analysis of Twitter Dataset through Random Forest Algorithm

Shubham Mishra[1], Vikas Pandey[2]
M. Tech. Student[1], Associate Professor (Dept. of Information Technology)[2]
Bhilai Institute of Technology, Durg, India [1,2]

Abstract

With the progression of web innovation and its development, there is a colossal volume of information present in the web for web clients and a great deal of information is produced as well. Web has turned into a stage for web based getting the hang of, trading thoughts and imparting insights. Long range interpersonal communication locales like Twitter, Facebook, Google+ are quickly picking up ubiquity as they enable individuals to share and express their perspectives about points, have dialog with various networks, or post messages over the world. There has been parcel of work in the field of opinion investigation of twitter information. Sentiment analysis examination is one of the unmistakable fields of data information mining that manages the distinguishing proof and investigation of wistful substance by and large accessible at online life. Twitter is one of such social Medias used by many users about some topics in the form of tweets. These tweets can be investigated to discover the perspectives and conclusions of the sentiments by using Random Forest Technique.

Keyword- online learning, sentiment analysis, data mining, social media

1. Introduction

The unrivalled increment in the acknowledgment just as entrance of web based life stages, for example, Facebook, Twitter, Google besides, and so on. In a day to day life, have changed the pattern of online communication of people. Formally, user’s online access was highly restricted to professional contents such as news agencies or corporations. In any case, they can flawlessly connect with one another in an increasingly simultaneous manner by making their very own substance inside a system of companions. Life has developed as an imperative stage of speaking to individuals' sentiments boosting the necessities of data mining information in the field of the conclusion investigation. In the slant examination, the crude information is the online content that is traded by clients through web based life. Twitter, which is one of such social Medias, has turned into the conspicuous source to trade the online content, giving an immense stage of supposition investigation. Twitter is a famous informal communication site that enables enlisted

clients to post short messages, likewise called tweets, up to 140 characters. Twitter database is one of the largest database having 200 million users who post 400 million messages/tweets in a day. At Twitter, clients regularly share their genuine belief on various subjects, for example, acknowledgment or dismissal of lawmakers and perspective about items, talk about current issues and offer their own life occasions. Be that as it may, clients post their tweets with less characters by used a short type of words and images, for example, emoticon. In this way, examination of these tweets can be used to discover solid perspectives and sentiments for any theme. Twitter information has just been use by various individuals to for see financial exchange forecast, film industry incomes for motion pictures. When study seriously and make daily decisions, it often see people's ideas. During the political election, they advised the forum of political discussion, reading consumers' reports during purchasing consumers asked friends and recommended dinner for dinner. Today, the Internet may find millions of people, from the latest gadgets to political philosophy. According to the latest Pew Internet and Civic participation survey, "only one in five people is posting material on political or social issues, or social networking sites for some citizenship or political participation.

2. Opinion Mining

Opinion analysis examination the otherwise called supposition mining, is a region of normal language handling (NLP) that constructs frameworks that endeavor to distinguish and separate feelings in content. Usually, in addition to opinion identification, these systems also extract expression attributes. Here is an example:

Polarity: off the chance that the speaker expresses a positive or negative sentiment.
Subject: what is being discussed.
Opinion holder: the individual, or element that communicates the conclusion.
Right now, feeling analysis examination is a theme of extraordinary intrigue and improvement since it has numerous pragmatic applications. As more information is publicly available and personally available on the Internet, there is a wealth of text on comment sites, forums, blogs, and social media.

3. Natural Language Processing (NLP)

The definition of Quirk's private status was used for story tracking (Wiebe, 2004). She defines the personal condition as a tuple (p, experience, attitude, purpose) and ties the experienced person's state to his/her attitude towards the purpose. In practice, models of simplified models are usually used. Here, only the goals of polarity and emotion are considered. In fact, many researchers broadly define emotion as a negative or positive perspective (Pang and Lee, 2002; Hu and Liu, 2005; Melville et al., 2009).

Sentiments also have some unique characteristics to distinguish them from other qualities that you want to track in your text. They usually want to categorize text by topic. It involves dealing with the entire topic classification. Emotional classification, on the other hand, usually has two categories (positive and negative), a range of polarity (such as a movie star rating) and even a range of opinion strengths (Pang and Lee, 2008). These classes span many topics, users, and various documents. Although it looks easier than standard text analysis because it can handle only a few classes, it is not.

4. Sentiment Analysis through NPL

Emotion analysis is a natural language processing and information extraction task that aims to get the writer's feelings expressed in positive or negative comments, questions, and wishes by analyzing a large number of documents. As a rule, sentiment analysis is gone for deciding the frame of mind of a speaker or author towards the general tonality of a point or archive. In recent years, the rapid increase in Internet usage and public opinion exchange has become the driving force of today's sentiment analysis. The Web is an enormous storehouse of organized and unstructured information. Breaking down this information to remove potential general supposition and feelings is a troublesome errand. Enthusiastic investigation might be archive based where feelings crosswise over reports are outlined as positive, negative or objective. It can be based on sentences, where one sentence in the text having emotions is classified. SA can be express based, and the expressions in the sentence are arranged by extremity. Emotion investigation distinguishes phrases with explicit emotion in the content. The author may talk about some objective facts or subjective opinions. They need to distinguish between the two. SA found a theme that emotions were aimed at. The content may contain numerous elements however, it is important to discover the substance to which the emotion is coordinated. It decides the extremity and the level of emotion. Emotions can be classified as objective (fact), positive (representing part of happiness, happiness or writer's satisfaction) or negative (representing sadness, frustration or disappointment for part of the author). Emotional scores can be given further based on their enthusiasm, negativity or impartiality. Areas of research are firmly identified with (or can be considered as a major aspect of) computational phonetics, normal language handling, and content data mining. Beginning with research on passionate status (brain research) and judgment (assessment hypothesis), this field endeavors to address new inquiries in other talk zones used new apparatuses given by data mining information and computational phonetics. Emotion examination has numerous names. Often referred to as subjective analysis, opinion mining, evaluation extraction, there are several links to emotional computing (computer recognition and emotional expression) (Pang and Lee, 2008). They study the subjective factors defined by. Man. As “lingual expression of private state in context” (Wiebe et al., 2004). These are usually single words, phrases, sentences. Although the entire document may be studied as emotional units (Turney and Littman, 2003; Agrawal et al., 2003), it is widely believed that emotions exist in smaller language units (Pang and Lee, 2008). These terms are used interchangeably as emotions and opinions often refer to the same concept.

5. Challenges for Sentiment Analysis

Emotion analysis methods are aimed at extracting positive and negative emotions from words from text, and classifying text as positive, negative or objective if it cannot find emotions from words. In this respect it can be regarded as a text classification task. In text classification, there are many classes that correspond to different topics. Also, in emotion analysis, there are only three main categories. Therefore, analysis of emotions seems easier than text classification, but it is not. The general issues can be summarized as follows:

5.1 Implicit Sentiment and Sarcasm

Even without verbal emotions, sentences can have understood emotions. Think about the accompanying model. Can someone sit in this movie? You should question the stability of the thinking of the writer who wrote this book. Both are negative sentences, but neither of the above two sentences explicitly has negative feelings. Therefore, identifying the meaning in SA is more important than grammar detection.

5.2 Domain Dependency

There are many words that differ in polarity from domain to domain. Consider the following example. This story is unpredictable. The maneuvering of the car is unpredictable. I go to read a book. In the first example, the emotions conveyed were positive, but the emotions conveyed in the second example were negative. The third example has positive emotions in the field of books, but has negative emotions (in which the director is asked to read a book) in the field of films.

5.3 Thwarted Expectations

Sometimes, the author deliberately sets the specific situation and toward the end can't help contradicting it. Think about the accompanying precedent. This motion picture should to be extraordinary. It sounds like a great plot, the actor is a first grade, and the cast is also very good, and Stallone is trying to deliver good work. But they cannot stand it. There are positive words in the direction, but the overall mood is negative for the last important sentence. And in traditional text classification, this is classified positively because the frequency of terms is more important than the existence of terms.

5.4 Pragmatics

It is significant to detect the pragmatics of the user opinion. It may change the mood completely. Consider the following example. He finished the Barca DESTROY Ac Milan finals and completely destroyed me. Capital letters can subtly express emotions. The first example shows positive emotions and the second example shows negative emotions. There are also many ways to express pragmatism.

5.5 World Knowledge

It is often necessary to incorporate world knowledge into the system to detect emotions. Consider the following example. He is Frankenstein. I have just finished Dr. Zivago for the first time, but I can only say that Russia is very bad. The first sentence represents negative emotions and the second sentence represents positive emotions.

5.6 Subjectivity Detection

This is to distinguish between stubborn texts and non-stubborn texts. This is used to improve system performance by including a subjective detection module to rule out objective facts. But this is usually difficult. Consider the following example. I hate love stories. I hate movies and “I hate stories”. The first example represents objective facts, and the second example represents an opinion on a particular movie.

5.7 Entity Identification

Text or sentences can have multiple entities. It is very important to find which entity the opinion is directed to. Consider the following example. Samsung beat Harry in a football game and is better than Nokia Ram. These precedents are useful for Samsung and Ram separately, yet negative for Nokia and Hari.

5.8 Negation Handling

Denial is a difficult task for SA. You can subtly express negation without explicitly using negative words. One way to explicitly handle negation in statements such as "I don't like movies" is to reverse the polarity (if not the opposite) of all the words that appear after the negation operator. But this does not apply to "I do not like performance, but I like this direction". So we also need to consider the scope of denial, which only spreads here. Therefore, what you can do is to change the polarity of all the words displayed after the negative word until another negative word is displayed. In any case, there are still issues. For example, in the sentence "I like not only performance but also direction", there is no reversal of polarity after "No" because "only" exists. Therefore, when designing an algorithm, you need to consider the combination of "no" and "only" words.

6. Classifier & Feature Extraction

6.1 Random Forest classifier is a tree-based classifier. It comprises of multiple classification trees that can be used to estimate the class label. For a given information, every tree vote for a specific category label and also the category label gaining the most votes are going to be appointed thereto data point. The error rate of this classifier depends on the correlation or association among any 2 trees within the forest additionally to the strength of definite or individual tree within the forest. so as to attenuate the error rate, the trees ought to be sturdy and also the degree of associativity ought to be as less as attainable. within the classifier tree, the interior nodes area unit portrayed because the options, the sides effort a node area unit portrayed as tests on the feature's weight, and also the leaves area unit portrayed as category classes. It performs classification preliminary from the foundation node and moves incrementally downward till a leaf node is detected. The document is then classified within the class that labels the leaf node.

6.2 TF-IDF Feature Extraction Method - TF-IDF technique relies on the frequency technique, it takes under consideration, not simply the incidence of a word in a very single sentence (or tweet) but within the entire corpus. TF-IDF works by penalizing the common words by assigning them lower weights whereas giving importance to words that ar rare within the entire corpus however seem in good numbers in few documents.

6.3 Classification Model- Twitter dataset is analyzed, we collected tweets, cleaned it, extracted feature from it and built a system based on Random Forest Classifier.

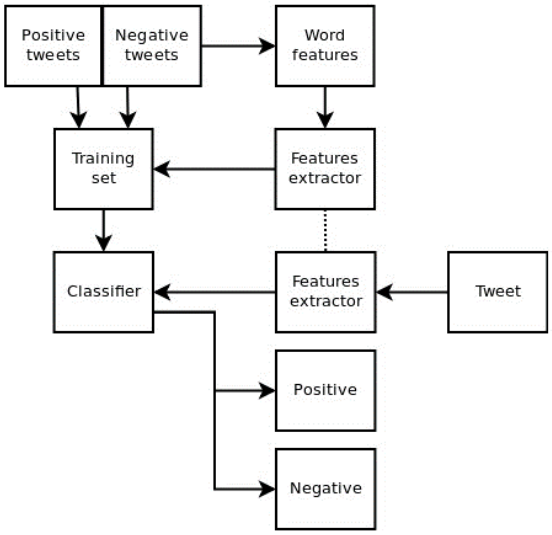

Fig 1: Classification Model.

7. Result and Discussion

7.1 Tweets Preprocessing and Cleaning

If the data is arranged in a structured format then it becomes easier to find the right information. This is initial step in which our data is preprocessed and cleaned. It is comparatively very easy to have meaningful insights from organized & clean data. The preprocessing is an essential step as it makes the raw text ready for mining. Tweets were analyzed and irrelevant words such as punctuation marks, fillers, colons, username, numbers etc were removed. Also, the cleaned tweets column is added to the training dataset.

7.2 Extracting Features from Cleaned Tweets

Now the dataset is cleaned, we need to analyze it. But first, it should be converted into features. There are various techniques to convert cleaned data into features. We will extract useful features from dataset using TF-IDF Feature Extraction Method.

7.3 Model Building: Sentiment Analysis

We are presently finished with all the pre-displaying stages required to get the information in the correct structure and shape. Now we will be building predictive models on the dataset using the TF-IDF Extracted Features.

7.4 Story Generation and Visualization from Tweets

In this area, we will investigate the cleaned tweets content. Investigating and imagining data mining information, regardless of whether its content or some other information, is a fundamental advance in picking up bits of knowledge.

7.5. Experiment Analysis

The datasets with no. of words in each tweet frequency is similar on both training and test dataset. Training Dataset has 31962 tweets, while test Dataset has 17197 tweets.

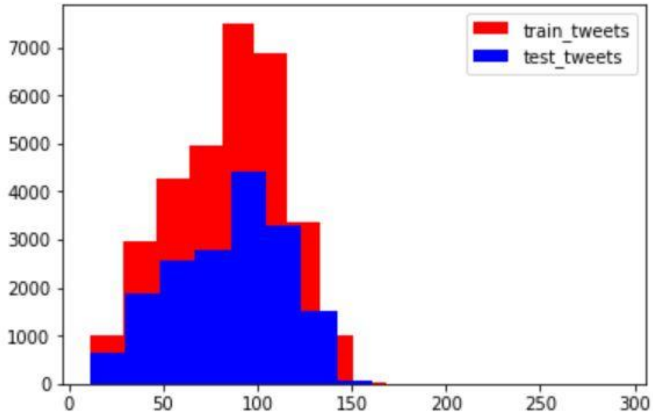

Fig 2: Train tweets vs. Test tweets (Dataset)

The model is trained with training datset and F1 score of 0.668493 is achieved. The model is then used on test dataset which found 799 Racist and 16398 non Racist tweets

As the below figure depicts that the comparative analysis has been found with the desire tweets with the non-desire tweets. This paper explores the nature of tweets from the given database. The dataset is prelabelled and is gathered from the twitter.

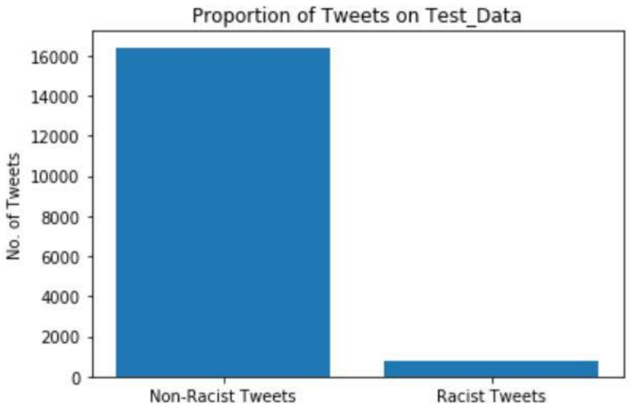

Fig 3: Tweets Comparison

9. Conclusion and Future Scope

This thesis focus on the study of sentiment analysis method to explore the key trend from the data sets. The program is using a machine-based learning approach which is more accurate for analyzing a sentiment than any other conventional method; together with natural language processing techniques will be used. We implemented Random Forest Classification Algorithm to analyze twitter dataset for classification of racist/non racist tweets from given dataset. Our model reached the efficiency of almost 67% that is the F1 score of 0.668493. The model resulted in classification of tweets on test raw data and found 799 Racist & 16398 non-Racist Tweets.

This work can be further extended by building twitter API application to fetch real time tweets using desired keyword. Also, the fetched tweets can be labelled using various machine learning techniques. Although Random Forest is best performing decision-tree based classification algorithm, above model can also be implemented using other techniques along with this technique in order to get more F1 score and better model effeciency. Next-generation opinion mining systems need a deeper bind between complete knowledge bases with reasoning methods inspired by human thought and psychology. This will lead to a better understanding of natural language opinions and will more efficiently bridge the gap between unstructured information in the form of human thoughts and structured data that can be analyzed and processed by a machine.

References

1. Agrawal, Rajagopalan, Srikant, and Xu (2003). Mining newsgroups using network arising from social behavior. *Twelfth international World Wide Web Conference*.
2. 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.
3. Cliche, M. (2017). BB_twtr at SemEval-2017 task 4: twitter sentiment analysis with CNNs and LSTMs. *arXiv preprint arXiv: 1704.06125*.
4. 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*.
5. Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2870-2879.
6. Kharde, V., & Sonawane, P. (2016). Sentiment analysis of twitter data: a survey of techniques. *arXiv preprint arXiv:1601.06971*.
7. 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*.
8. 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*.
9. Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 764-779.
10. 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.
11. Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135.
12. Singh, T., & Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. *Procedia Computer Science*, *89*, 549-554.
13. 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.
14. Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30:277–308.