Novel Sentiment Analysis using Twitter

Athare Sharayu Shivaji
Student, M.E. Comp. Engg
V.A.C.O.E. Ahmednagar

Rathod Vijay Uttam
Assistant Professor,
Department of Comp. Engg
V.A.C.O.E Ahmednagar

ABSTRACT

Sentiment analysis is a natural language processing tool that is useful for monitoring and accessing information from Web applications, as it is a treasure of public opinion about numerous issues without requiring to find and asses authentication of the document. With the rapid growth of social networks and microblogging websites, communication between people from different countries, cultural and psychological and physical backgrounds became more direct, resulting in more and more conflicts about thoughts and speech used between these people. The meaning of Hate speech can be explained and put into context as the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or beliefs and religion, political parties they resemble. As in India a lot of hate speech generated posts are placed on social networking sites. So to block and catch the hate speech generated posts to avoid necessary conflicts. So we thought of introducing a technique using Twitter as our data source. As twitter is generally and widely used by millions of Indians it is great source of raw data that can be monitored and made sharable without conflict. The idea came up of creating and analyzing datasets that can be used for machine Learning. By using political leaders as the study topics e.g. Narendra Modi, Rahul Gandhi etc. It is used to analyze the negative and positive hate speech tweets generated on them. Machine Learning algorithms are used to analyze the tweets and find the correct meaning behind it whether it is offensive or not. In this project we are various techniques such as Stop Words, Lexicon Analysis, Datasets and Machine Learning are used to analyze tweets and find out the sentiments behind it. Apache Spark based parallel processing technique is also used to access only the latest tweets and not the old ones which are already being analyzed.

General Terms

Hate speech

Keywords

Twitter, Sentiment Analysis, Stop words, Lexicon Analysis, Machine Learning, Apache Spark.

1. INTRODUCTION

Social networking sites(SNS) and microblogging websites are attracting internet users more than any other kind of website presently used by the users of the web. Services offered by Twitter, Facebook and Instagram are more and more popular among people from different countries, backgrounds, cultures and interests and political parties. Their contents are rapidly growing, constituting a very interesting example of the so-called big data with millions and millions sharing their data each minute. Big data have been attracting the attention of researcher and organizations, who have been interested in the automatic analysis of people's opinions and the structure/distribution of users in the networks, etc. and use them for a proper and beneficial use. While these SNS offer an open space for people to discuss and share thoughts and opinions, their nature. But as huge number of posts, comments and messages exchanged makes it almost impossible to control their content and hazards arousing from it. Furthermore, given the different backgrounds, cultures, beliefs and ethnicity, many people tend to use and aggressive and hateful language when discussing with people who do not share the same backgrounds and ethnicity. Many US papers reported that 481 hate crimes with an anti-Islamic motive occurred in the year that following 9/11, 58% of them were perpetrated within two weeks after the event after 9/11[5]. However, nowadays, with the rapid growth of SNS, more conflicts are taking place, following each big event or other every day. So to avoid such conflicts a novel approach which will detect offensive speech in a tweet and help in avoiding it is proposed.

2. MOTIVATION

Sentiment analysis using tweets is a hot topic. But as tweets are in large numbers generated daily from all around the world it has been not been easy to track them. Many are offensive tweets and may create a conflict between two users[14].To solve this problem, sentiment analysis using tweets by creating a training dataset of our own is suggested in this work. The readymade machine learning datasets are old and not resemble the current sentiments of fast evolving worlds. As tweets are generated in large numbers it uses big data techniques. These techniques are not easy to access. So to catch the latest tweets from a large twitter network we thought of using Apache Spark as a parallel processing tool to catch latest tweets and not the old ones. By doing this we find out the current trends and current topics that are discussed in twitter.

3. LITERATURE SURVEY

This topic of previous studies describes the fundamentals of tweet analysis. It helps in understanding and evaluating various ideas put forward by various technical papers published by various publishers.

3.1 A Real-world Dataset for Weakly Supervised Cross-Media Retrieval, IEEE, 2017.

Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang explains as their paper contributes a new large-scale dataset for weakly supervised cross-media retrieval and analysis , named Twitter100k. Current datasets, such as Wikipedia, NUS Wide and Flickr30k, have two major limitations and are upto date. First, these datasets are lacking in content diversity, i.e., only some pre-defined classes are covered where the current classes are excluded. Second, texts in these datasets are written in well-organized language, leading to inconsistency with realistic applications and does not cover many countries. To overcome these drawbacks, the proposed Twitter100k dataset is characterized by two aspects: 1) it has 100,000 image-text pairs randomly downloaded from Twitter and thus has no constraint in the image categories; 2) text in Twitter100k is written in informal language by the users and not hard to understand. Since strongly supervised methods creates the class labels that may be missing in practice, here our technique focuses on weakly supervised learning for cross-media retrieval, in which only text image pairs are exploited during training and more labels can be added to it. We thoroughly analyze the performance of four learning methods and three variants with various text features on Wikipedia, Flickr30k and Twitter100k. As a minor contribution, a deep neural network to learn cross-modal embedding for Twitter100k is designed which works differently from Machine Learning. Inspired by the characteristic of Twitter100k, a method to integrate Optical Character Recognition (OCR) into cross-media retrieval for recognizing words and numbers is proposed.

3.2 Sentiment Analysis in TripAdvisor, IEEE, 2017.

Ana Valdivia, M. Victoria Luzón, and Francisco Herrera studies TripAdvisor SNS based Sentiment Analysis. According to Wikipedia, TripAdvisor is an American travel SNS providing reviews from travelers about their experiences in hotels, restaurants, and monuments. Stephen Kaufer and Langley Steinert, along with others, founded TripAdvisor in February 2000 as a site listing information from guidebooks, newspapers, and magazines for travelers that can be used by others. After that, the website turned to user-generated content. It has since become the largest travel community, reaching 390 million unique visitors each month and listing 465 million reviews and opinions about more than 7 million accommodations, restaurants, and attractions in 49 markets worldwide and becoming a widely used SNS for travelling purpose. So it can be used as a text source, storing numerous real time reviews of tourist businesses around the world. Sentiment analysis extracts insights from this data for analysis purpose. Sentiment classification, the best-known sentiment analysis task, aims to detect sentiments within a document, a sentence, or an aspect from TripAdvisor SNS.

3.3 Hate Speech on Twitter, IEEE, 2018

Hajime Watanabe, Mondher Bouazizi and Tomoaki Ohtsuki published a paper which studies hate speech sentiments in tweets. It says that with the enormous growth of social networks and microblogging websites, communication between people from different countries, cultural and psychological backgrounds became more direct, resulting in more and more “cyber” conflicts between these people interacting on theses SNS. Thus hate speech is used more and more, to the point where it became a serious problem invading these open spaces which increases misuse[3]. Hate speech can be explained as the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or their beliefs and religion, etc. While most of the online SNS forbid the use of hate speech, the size of these networks and websites makes it almost impossible to control all of their content. Therefore, it gives rise to the necessity to detect such speech automatically and filter any content that presents hateful language or language inciting to hatred. Authors proposed an approach to detect hate expressions on Twitter by analyzing them. Their approach is based on unigrams and patterns that are automatically collected from the training set before machine learning. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm for analysis.

3.4 Social Media’s Role in Democracy, IEEE, 2018.

Wendy Hall, Ramine Tinati, and Will Jennings published a paper which studies TRUMP and BREXIT tweets for sentiment analysis. It mainly tells that to understand the role of social media during political events, Twitter datasets were used to describe in and a collection of analytical methods to help describe the structure and content of the network as well as the interactions between humans and their importance to analyze the tweets and use them for various purposes. The analysis focused on the three main areas in question: the temporal evolution of the network, its structure, and the topics and sentiment of content within it and tweets will be used for this purpose. Brexit—the UK’s referendum on EU membership—was Europe’s defining political event of 2016, and it was without doubt the most significant political event in the UK in half a century. The official (and unofficial) campaigns, politicians, and citizens engaged heavily in discussions on SNS platforms to communicate Brexit-related information and arguments, gain attention, and influence voters. In contrast to the Brexit campaign, which was a one-off event, the 2016 US presidential election is part of a long-established political cycle which follows for each presidential election. For several US presidential campaigns, SNS has been a prominent venue for discussion and a focus of engagement for candidates and news media. In the lead up to the election, the discourse was dominated and used by the selected candidates of the Republican and Democratic parties, Donald Trump and Hillary Clinton, respectively. Thus Sentiment can be used for a meaning purpose.

4. PROPOSED SYSTEM

The Sentiment Analysis model main aim is to provide solution to analyze lakhs of tweets that are generated every second on twitter. It is low cost and efficient system. It includes Apache Spark, Twitter, Preprocessing, Machine learning etc.

In proposed system first the tweets are accessed using Apache spark and Twitter4j API. Then they are preprocessed and then machine learning is applied to it to get the sentiment.

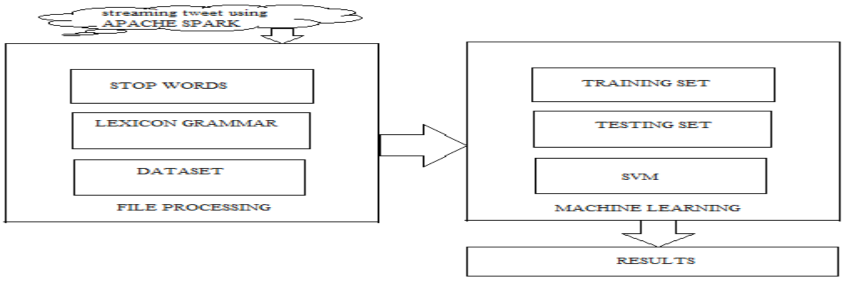

Fig 1: System Architecture

4.1 Tweet access using Apache Spark Module

This module first initializes Apache spark over a local IP address, then authentication is provided to twitter by creating a developer account and get authentication details for success. The apache spark will access clusters of twitter and access only the latest tweets.

4.2 Preprocessing Module

In this module the tweets will be accessed from text file and then the stop words removal is generated. The stop words which are not good for text mining are matched with an array of stop words and unnecessary words are removed from the tweets. The preprocessed tweets are then stored in a separate text file.

4.3 Lexicon Grammar Module

In this module the preprocessed text file is accessed and AFFIN library is applied to it. It returns adjective, noun and pronoun of a word. We only access pronouns and shorten the tweet further. The shorten tweets are again stored in a separate text file.

4.4 Machine Learning Module

In this module first a training dataset is designed with two classes positive and negative. Then a tweet is accessed and a test dataset is generated. Then an instance of SVM classifier is generated and training and testing dataset is applied to it. It returns the result in the form of parameters such as tp rate, fp rate, precision and recall. We take into account precision

if it is greater than 0.5 then the tweet comes under positive sentiment and if it is smaller than 0.5 the tweet comes under negative sentiment.

5. EXPERIMENTAL RESULTS

The extraction of features and optimization of parameters is done. The classification is done using the toolkit weka. Weka presents a variety of of classifiers organized into groups based on the type of the algorithm(e.g., decision tree-based, rule-based,etc)[3].Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to the functions. To evaluate the performance of classification, different key performance indicators(KPIs) are used which are the percentage of true positives, the precision, the recall.

6. CONCLUSION

In this project, novel sentiment analysis approach using TWITTER and APACHE SPARK together is developed. The The basic idea of the project is to use distributed computing in training and testing the machine learning classification. The various predictions are assembled by machine learning algorithms together and view the results in three classes such as positive, negative and neutral according to the predictions returned by the system.

7. ACKNOWLEDGMENTS

We would like to give our sincere thanks to all the staff of V.A.C.O.E Ahmednagar who helped us directly & indirectly.

8. REFERENCES

[1] Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang,Twitter100k: A Real-world Dataset for Weakly Supervised Cross-Media Retrieval, IEEE, 2017.

[2] Ana Valdivia, M. Victoria Luzón, and Francisco Herrera, Sentiment Analysis in TripAdvisor, IEEE, 2017.

[3] Hajime Watanabe, Mondher Bouazizi and Tomoaki Ohtsuki, Hate Speech on Twitter, IEEE, 2018.

[4] Wendy Hall, Ramine Tinati, and Will Jennings, From Brexit to Trump: Social Media’s Role in Democracy, IEEE, 2018.

[5] B. Pang and L. Lee, Opinion mining and sentiment analysis, Foundations Trends Inf. Retriev., vol. 2, no. 12, pp. 1135, 2008.

[6] B. Liu, Sentiment analysis and opinion mining, Synth. Lectures HumanLang. Technol., vol. 5, no. 1, pp. 1167, 2012.

[7] C. Havasi, E. Cambria, B. Schuller, B. Liu, and H. Wang, Knowledgebased approaches to concept-level sentiment analysis, IEEE Intell. Syst., vol. 28, no. 2, pp. 001214, Mar.-Apr. 2013.

[8] C. D. Manning and H. Schtze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[9] P. D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in Proc. ACL, 2002, pp. 417424.

[10] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, Lexiconbased methods for sentiment analysis, Comput. linguist., vol. 37, no. 2, pp. 267307, 2011.

[11] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in Proc. EMNLP, 2002, pp. 7986.

[12] J. Zhao, L. Dong, J. Wu, and K. Xu, Moodlens: An tweets, in SIGKDD, 2012 emoticon-based sentiment analysis system for chinese is used .

[13] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, Learning word vectors for sentiment analysis, in Proc. ACL, 2011.

[14] G. Paltoglou and M. Thelwall, A study of information retrieval weighting schemes for sentiment analysis, in Proc. ACL, 2010, pp. 13861395.