Sentiment Analysis of Live Twitter Data using Apache Spark

Vemula Sai Saketh[1], Yashvanth Kumar Guntupalli[2], Devashish S Vaishnav[3]

[1-3]PES University, Bangalore, India

---***---

Abstract - *Sentiment Analysis is the method of utilizing content analytics to mine different sources of information for conclusions. Frequently, Sentiment Analysis is done on the information that's collected from the Web and from different social media platforms. Politicians and governments frequently utilize opinion examination to get how the individuals feel about themselves and their approaches. With the advent of social media, data is captured from different sources, such as mobile devices and web browsers, and it is stored in various data formats. Because the social media content is unstructured with respect to conventional capacity systems (such as RDBMS, Relational Database Management System), we require devices that can process and analyze this disparate data. However, big data technology is made to handle the different sources and different formats of the structured and unstructured data. In this paper, we perform sentiment analysis with the help of Apache Spark framework, which is an open source distributed data processing platform which utilizes distributed memory abstraction. The effectiveness of our proposed approach is proved against other approaches achieving better classification results when using the Multinomial Naïve Bayes classification algorithm.*

Keywords: *Sentiment Analysis, Apache Spark, Big Data, Multinomial Naive Bayes, Machine Learning Models, API.*

1. INTRODUCTION

Nowadays micro blogging applications like Twitter, Facebook etc. have been trending since they help express peoples attitude or concern towards a certain topic which is usually depicted through a hashtag and a lot of other factors like emoticons, emojis etc. The apprehension and assortment of emotions within text data using text analysis techniques are called Sentiment Analysis. A lot of unstructured data is retrieved everyday from various social media sites and micro blogging sites.

This data is hard to read through end to end and manually assigning an emotion. Sentiment Analysis eradicates this problem and automates this process. This niche technology proved to be useful in many ways in the real world. Suppose there is a new product release by a certain company this technique is used in the feedback forum so as to get the review of the product whether it is positive or negative. Same data can be used to find customers who have a negative opinion on the product and address their concerns to get an overall positive image on the product and hence increase the sales. Another use case is, if one is thinking of investing in a particular company, future trends can be predicted in real time by tracking and studying consumer behaviour by using this type of computational linguistics. Since social media sites handle huge amounts of data on a regular basis a powerful tool is required to manage, handle and retrieve it. Big Data would aid in achieving this. Big Data would provide us multiple applications such as Hadoop, Apache Spark, Apache Flume and distributed data storages like Hadoop Distributed File System. Apache Spark is a data processing framework that can quickly perform processing tasks on very large data sets, and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools. These two qualities are key to the worlds of big data and machine learning, which require the marshalling of massive computing power to crunch through large data stores. Spark also takes some of the programming burdens of these tasks off the shoulders of developers with an easy-to-use API that abstracts away much of the grunt work of distributed computing and big data processing. Apache spark has a library called Spark Streaming using which we achieve our desired functionality. In this paper our goal is to use Apache Spark framework to get the live stream of tweets from Twitter using Twitter Developer API Account and predict their sentiment using Multinomial Naive Bayes classification technique.

Our paper aims to make a Sentiment Analysis application available as a service to everyone. We will use Apache Spark to extract live tweets from twitter using their developer API, analyze the sentiment and display it on user interface. This application will be hosted on AWS so it can be accessed from anywhere. We have included two services of which one is used to classify the real time tweets and the other is used to classify movie reviews which are entered by the users.

2. RELATED WORKS

Sentiment Analysis has been a hot topic since the past couple of years. There was edge-cutting research done on it and a lot is yet to be uncovered on using this technology to its full potential. In [1], a list of various Sentiment Analysis algorithms were described. They were categorised into Machine learning based and lexicon based approaches for which the advantages and disadvantages were provided. [2] has utilized the combination of Naive Bayes algorithm and SentiWordNet to introduce a novel method through which a greater accuracy could be achieved. Applying Naive Bayes with SentiWordNet in the classification of text would result in better accuracy. In the U.S. voting Presidential Election Cycle held in 2012, [3] built a real time Twitter Sentiment Analysis system using a training dataset of 17000 related tweets. In [4] using Apache Spark platform, a sentiment learning model was built to run on it. Using the suggested algorithm the pre-processing of hashtags and emotions within a tweet were exploited, after which sentiment types were classified using parallel processing methods. Using spark in Big Data an efficient sentiment prediction technique was recommended in [5]. High levels of scalability in relation to accuracy and time were achieved through this novel technique. The processing time indicates less variance with the growth of data volume. Preprocessing of data to ignore noise in the data was suggested in [6], and hence implemented sentiment analysis with a great number of tweets by making use of Hadoop framework. With accelerated data thriving everyday it is important to refine only what is in the interest of the user, so there is a need to automate this process. Additionally, there is also a need to address problems like overlooking grammatical errors by using TF-IDF which has been proposed in [7].

Approximately 10,000 tweets were classified into positive, negative and neutral by applying machine learning and pattern recognition techniques in [8] which gained an accuracy of 70%. Authors of [9] proposed that considering syntactic properties could play an important role in sentiment classification. Into a naive bag of words multiple sentiment features based on syntax trees have been introduced and machine learning methods Naive Bayes and Support vector machines were trained on movie dataset to provide accurate solutions. [10] proposes that combining supervised and unsupervised machine learning models would lead to strong performance on text data. The authors tested this on tweets extracted for comparing McDonalds and KFC to show which one was better. In [11] the author has used Possibilistic Fuzzy C-Means with SVM to achieve high levels of accuracy on movie tweets and worked on upto 3-grams. Building a domain oriented approach by considering domain independent and domain specific lexicons in the area of smartphone brands have shown a rise of around 2 points on an average over the unigram baseline in [12]. Authors of [13] have proven that Spark system has rapid computational time compared to Hadoop Mapreduce while working on corpus based sentiment analysis. To calculate the sentiment of a given sentence or paragraph is broken into words which are categorised into positive and negative words using an existing database of words and based on the count of positive and negative words the polarity was analysed in [14]. [15] proposes that there is a significant effect on the sentiment analysis by considering domain specific texts and hence the authors have performed analysis on electronic products. The potency of Apache spark while handling huge amounts of data for performing machine learning tasks like classification, regression and dimension reduction has been highlighted well in [16]. Even with limited computational power, typically in a standard single computer Apache spark's ability to classify expertly has experimentally proven in [17]. [18] talks about the importance of how performing sentiment analysis on customer opinion in real time can tremendously affect one's decision making process for a business domain. The strong effect of taking emoticons present in the text while classifying sentiment was proven to have higher impact in [19]. Authors of [20] while performing sentiment analysis on Greek data have found that there is a significant effect of considering emoticons into classification to get better results.

3. MACHINE LEARNING TECHNIQUES

Machine learning has numerous uses in today's world of technology. This state-of-the-art technology is not limited to one domain. The sharing and tagging of friends on Social Media is done using Image Processing which is a Machine Learning technique. Optical character recognition which is also a Machine Learning application widely used for transforming Images of typed or handwritten text into machine encoded text. Not only these there are many common life applications like predicting traffic while commuting, Virtual personal assistants like Siri, Alexa, Google, Email Spam and Malware Filtering, Online customer support and point of interest Sentiment Analysis. It is important to understand the impact of Machine Learning since it is a continuously developing field.

3.1 Multinomial Naive Bayes

Generally Naive Bayes assumes independence in the model, rather than the particular distribution of each feature. In layman's terms Naive Bayes assumes that the features it uses are conditionally independent of each other. So if one has to calculate the probability of observing features through , given some class c, under the Naive Bayes assumption the following holds:

p(f1,........fn|c) = i=1np(fi|c)

This assumption of independent features has resulted very well when performed on complex tasks where it was stated that strong independent assumptions are false. The posterior probability of the above equation is as follows:

p(c|f1,........fn) p(c)p(f1|c)......p(fn|c)

Which is simpler to work with. The term Multinomial Naive Bayes simply lets us know that each is a multinomial distribution, rather than some other distribution. This works well for data which can easily be turned into counts, such as wordcounts in text.

3.2 Linear SVC

The aim of a Linear SVC (Support Vector Classifier) is to fit to the data you contribute, resulting in a "best fit" hyperplane that partitions, or segregates, your data. From there, after capturing the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this definitive algorithm rather convenient for our uses, though you can use this for enumerable situations. Some prevalent applications of this approach include Face detection, text and hypertext detection, classification of images, bioinformatics, protein fold and remote homology detection, handwriting recognition, and generalized predictive control.

3.3 Bernoulli Naive Bayes

BernoulliNB implements the Naive Bayes training and classification algorithms for input that is scattered bestow to multivariate Bernoulli distributions; i.e., there may be numerous features but each one is simulated to be a binary-valued (Bernoulli, boolean) variable. Consequently, this class requires samples to be expressed as binary-valued feature vectors; if delivered any other kind of input, a BernoulliNB instance may binarise its input. The decision rule for Bernoulli naive Bayes varies from multinomial NB's rule in that it explicitly penalizes the

non-occurrence of a feature i that is an indicator for classy, where the multinomial variant would openly neglect a non-occurring feature.In the case of text classification, word occurrence vectors (rather than word count vectors) may be used to train and use this classifier. BernoulliNB might achieve better on some datasets, exclusively those with reduced documents.

p(x) = P[X = x] ={ p x = 1q = p-1 x = 0
P(xi | y) = P(i | y)xi + (1- P(i | y))(1 - xi)

3.4 Sentiment Analysis

Sentiment Analysis is the method of utilizing content analytics to mine different sources of information for conclusions. Frequently, Sentiment Analysis is done on the information that's collected from the Web and from different social media platforms. Politicians and governments frequently utilize opinion examination to get how the individuals feel about themselves and their approaches. With the advent of social media, data is captured from different sources, such as mobile devices and web browsers, and it is stored in various data formats. Because the social media content is unstructured with respect to conventional capacity systems (such as RDBMS, Relational Database Management System), we require devices that can process and analyze this disparate data. However, big data technology is made to handle the different sources and different formats of the structured and unstructured data.

4. BIG DATA & CLOUD COMPUTING

A huge collection of data growing day-to-day exponentially with time is coined as big data. This data could be of structured or unstructured type. Analysing and extracting key factors from the big data can help a company perform better. The characteristics of Big data are Volume (Huge amounts of data), Velocity (day-to-day exponential growth of data), Variety (Different types of data like images, videos, text, audio etc.) Since Big data is usually large amounts of data and is difficult to store locally, so we use a technology termed cloud. Cloud computing is basically on demand supply of computing services - applications and storage. Storing and accessing data is made easy with the help of this technology. This technology avoids the up-front of owning and maintaining their own IT infrastructure.

4.1 Apache Spark

To handle the huge amounts of data we need a platform to process the same efficiently, Apache Spark helps us in achieving this. It not only increases the usability but also divides the tasks among multiple computers which makes it the ideal tool for handling Big Data. Spark streaming is a library which is used extensively by all. It helps processing of real time data from various sources and this processed data can be pushed out to file systems, databases. In the real world it is typically used to stream the tweets from twitter using their developer account.

4.2 AWS EC2

The cloud services provider Amazon's widely used Elastic Computing provides infrastructure for users to deploy their applications. One of the key features of this service is that when the running application needs more resources than allocated EC2 automatically allocates them. Coming to the security of data users can control which instances remain private and which are to be exposed to the internet. EC2 leverages Amazon Virtual Private Cloud (VPC) for security, and businesses can connect their secure IT infrastructure to resources in VPC.

5. PROPOSED SOLUTION

Using twitter tweet data we are making a sentiment analysis model which predicts the sentiment of a given tweet/text. The sentiment analysis application is made available as a web application where users can login and check the sentiment of tweets they are interested in. The APIs created for the sentiment analysis application can be accessed globally since it has been hosted on cloud. Our solution takes emoticons into consideration since a lot of tweets had emoticons involved in them and these play an important role in identifying sentiment. This sentiment analysis application is made available as a service which can be invoked from anywhere using the API. The web application provides a dashboard where users can test the sentiment of the tweet or response. Spark handles live streaming data. The future scope of our project is to make this Application available to the public and also companies which might require some sentiment analysis. It is known that Spark can handle huge amounts of data for example yelp's add platform handles millions of add requests everyday.

6. DATA SET

Sentiment140: This dataset consists of around 1.6 million tweets. Each tweet is tagged with two sentiments where 0 and 4 represents negative and positive respectively. It consists of 6 fields namely target, ids, date, flag, user and text.

7. EXPERIMENTATION AND RESULTS

The architecture of the project has been proven successful in terms of the end goals. Multiple models like Multinomial NB, Linear SVC, Logistic Regression, Bernoulli NB and Gradient Boosting algorithms were used for analysing the sentiment of tweets. We used a dedicated set of training and testing data sets to do that. After this we have created a couple of API to expose the previously implemented models to the outside world. These REST API's will be the POST methods. Input is taken from the user which is fetched to the models built using these API and the sentiment analysed is displayed. Using the Apache spark's streaming library we have written a scala code which takes the twitter developer account credentials to fetch live tweets. The sentiment of the tweets is then analysed using our model and the results are stored and displayed on an user interface. The interface provides multiple features like providing filters for selecting tweets containing only those hashtags of their interest. It also provides a count of positive and negative tweets analysed.

These stored tweets are further passed into the training data so as to increase the accuracy of the model. The following results were achieved:

Table 1: Results

Model	Accuracy
Bernoulli Naive Bayes	75.2
Linear SVC	79.7
Multinomial Naive Bayes	81.3

We could use Apache spark to get the live stream of tweets and analyse their sentiment. A web application hosted on cloud which provides an interface to users to analyze tweets, test their tweets and also rate and review movies.

8. CONCLUSIONS

An efficient Sentiment Analysis system has been proposed in this project using Multinomial Naïve Bayes and the Linear SVC algorithms to perform the estimation of sentiment tweets and movie reviews. The accuracy found seems to be promising and the web service component gives the advantages to the third party applications to readily integrate the solution into their apps without having to do a lot of coding. Using Apache Spark to stream the live tweets helps organisations analyse tweets in real time and can be used for their betterment. The Web interface created helps users get more insights of tweets they are interested in and also the MovieReviewAnalysis component shows the ratings of movies, allows users to add new movies, read and write reviews for

REFERENCES

[1] Medhat, Walaa, Ahmed Hassan and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5 (2014): 1093-1113.

[2] A. Goel, J. Gautam and S. Kumar, "Real time sentiment analysis of tweets using Naive Bayes," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2016, pp. 257-261, doi: 10.1109/NGCT.2016.7877424.

[3] Wang, Hao & Can, Dogan & Kazemzadeh, Abe & Bar, François & Narayanan, Shrikanth. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle. 115-120.

[4] Samar Al-Saqqa, Ghazi Al-Naymat, Arafat Awajan, A Large-Scale Sentiment Data Classification for Online Reviews Under Apache Spark, Procedia Computer Science, Volume 141, 2018, Pages 183-189, ISSN 1877-0509.

[5] V. J. Nirmal and D. I. G. Amalarethinam, "Real-Time Sentiment Prediction on Streaming Social Network Data Using In-Memory Processing," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, 2017, pp. 69-72, doi: 10.1109/WCCCT.2016.26.

[6] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.

[7] C. Liu, Y. Sheng, Z. Wei and Y. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Lanzhou, 2018, pp. 218-222, doi: 10.1109/IRCE.2018.8492945.

[8] A. I. Baqapuri, S. Saleh, M. U. Ilyas, M. M. Khan and A. M. Qamar, "Sentiment classification of tweets using hierarchical classification," 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016, pp. 1-7, doi: 10.1109/ICC.2016.7511391.

[9] H. Zou, X. Tang, B. Xie and B. Liu, "Sentiment Classification Using Machine Learning Techniques with Syntax Features," 2015 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, 2015, pp. 175-179, doi: 10.1109/CSCI.2015.44.

[10] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[11] R. D. Desai, "Sentiment Analysis of Twitter Data," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 114-117, doi: 10.1109/ICCONS.2018.8662942.

[12] M. Venugopalan and D. Gupta, "Exploring sentiment analysis on twitter data," 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, 2015, pp. 241-247, doi: 10.1109/IC3.2015.7346686.

[13] J. Ranganathan, A. S. Irudayaraj and A. A. Tzacheva, "Action Rules for Sentiment Analysis on Twitter Data Using Spark," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, 2017, pp. 51-60, doi: 10.1109/ICDMW.2017.14.

[14] B. Nandi, M. Ghanti and S. Paul, "Text based sentiment analysis," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 9-13, doi: 10.1109/ICICI.2017.8365326.

[15] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-5, doi: 10.1109/ICCCNT.2013.6726818.

[16] M. Assefi, E. Behravesh, G. Liu and A. P. Tafti, "Big data machine learning using apache spark MLlib," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 3492-3498, doi: 10.1109/BigData.2017.8258338.

[17] D. Andrešić, P. Šaloun and I. Anagnostopoulos, "Efficient big data analysis on a single machine using apache spark and self-organizing map libraries," 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Bratislava, 2017, pp. 1-5, doi: 10.1109/SMAP.2017.8022657.

[18] S. Chaturvedi, V. Mishra and N. Mishra, "Sentiment analysis using machine learning for business intelligence," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2162-2166, doi: 10.1109/ICPCSI.2017.8392100.

[19] H. Wang and J. A. Castanon, "Sentiment expression via emoticons on social media," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 2404-2408, doi: 10.1109/BigData.2015.7364034.

[20] G. S. Solakidis, K. N. Vavliakis and P. A. Mitkas, "Multilingual Sentiment Analysis Using Emoticons and Keywords," 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, 2014, pp. 102-109, doi: 10.1109/WI-IAT.2014.86.