Journal of Natural Sciences and Engineering, Vol. 3, (2020)
DOI number: 10.14706/JONSAE2020114

Sentiment Analysis on Twitter Data using Big Data

Obada Almonajed, Samed Jukić
[1]International Burch University, Sarajevo, Bosnia and Herzegovina
almonajed.obada@ibu.edu.ba
samed.jukic@ibu.edu.ba

Abstract –With the increasing number of users and data on the Internet, especially social media sites, sentiment analysis topic became one of the important and essential fields for most. Collection of people's feelings and sentiment and classifying the data attracted most businesses and companies. Recently, twitter sentiment analysis has attracted much attention, because of Twitter's growth and popularity. The solution for handling enormous amounts of data from social media is a new term called Big data. Big data is not just for having a large amount of data, but also the importance of processing and the usage of the data. In this paper, we collect live data from Twitter using Apache Spark; and apply machine learning algorithms provided by Apache Spark machine learning library for classification of each Twitter message. Naive Bayes and Logistic Regression are used for testing the model. Naive Bayes algorithm gave better results, where it has an average accuracy around 75%, while the Logistic Regression algorithm was around 69%.

Keywords–big data, sentiment analysis, twitter, apache spark, social media, machine learning.

1. Introduction

Social media, one of the best things about it is in its name; social. It connects various people across the world by sharing information to them and receiving information from them. The main purpose of social media is to connect people and allow them to share thoughts and opinions. It allows also to read the news, watch videos, read stories, view and share photos. Social media is becoming an integral part of our lives. It is a way of connecting and building a relationship with others. It allows you to hear what people say and to respond. The most popular platforms are Facebook, Twitter, YouTube, Instagram, Snapchat.

Since social media allows people to connect those days social media are very important for businesses. It takes advantage of social media to increase brand exposure and customer reach. Publishing to social media is very simple. For example, a company can create a page on Facebook, and post new products, sales announcements, market brands, and products as images or text or video. No matter the size of the business, it is important to recognize the value and trend for better understanding and utilizing the platform.

People can talk about your business without your knowledge. So, as a company, it is important to know and monitor social media conversations about the brand. Based on reviews, the company can always adjust the present market situation and satisfy customers in a better way. In order to identify the text written by

your customers, a sentiment analysis tool is used. Sentiment analysis or opinion mining is used to determine the emotional tone of message or text. The main usage of this tool is to understand how people feel and think about something. The tool is very useful for companies and can affect decision making.Using machine learning, companies can analyze the content on social media to see the meaning behind the messages.

An Enormous number of people across the world use social media. In order to gain such data, store, and process, we will use Big data. Big data is not only for storing a large amount of data but the ability to analyze. Big data allows us to get and analyze real-time data from social media. For this paper, one of the fastest big data platforms Apache Spark will be used. Compared with Hadoop, it can be faster up to one hundred times[1]. Apache Spark framework provides native bindings for Java, Python, Scala, Machine Learning, and support SQL. The purpose of the paper is to collect data from Twitter and determine and classify the feeling of the user into positive or negative using machine learning and Apache Spark.

2. Literature Review

Pang et al. [2], in the paper, they came out that unigram is a better model over others. Regardless of whether there is no large difference between unigram precision and mix of unigrams and bigrams precision, where the precision using unigrams has 82.9% and precision using the mix of unigrams and biagrams is 82.7%; both predicted with SVM algorithm. However, Dave et al. [3] have inverse results, where bigrams gave preferable precision over unigrams utilizing SVM and Baseline algorithms. SVM brings about 87.2% precision for the first test and 85.8% precision for the second test for bigrams.

Pak et al. [4] gathered around 300.000 various tweets for Twitter. The tweet can be classified into three classes, positive, negative, or neutral. They thought about that, the emoji in the message represents the actual sentiment of the text. Thus, if ':(' emoji is included in the message, regardless of what is the content: the message has negative sentiment. Likewise, if a tweet has ':)', the message is considered as negative sentiment. For learning algorithms, they utilized multinomial Naïve Bayes, SVM and Conditional random fields, yet Naïve Bayes indicated the best outcomes. To make the precision of the classifier better, they removed some n-grams, since it isn't showing any sentiment.

Authors of the paper [5], have researched the usage of Apache Flume and Apache Hive which is built on top Hadoop for analyzing Twitter data. In the research[6], the authors wrote and discussed a recommendation system that provides a summary of users' feedback, comments, and reviews about different subjects using the Hadoop framework. Similarly, the authors of the researches [7], built a recommendation system that recommends services. The researchers of the paper[8], build a Hadoop framework for determining and analyzing the customers' feedback toward a product from social networks, that framework extracts and analyzes the feedback of social user relationship management.

Go et al. [9] broke down Twitter suppositions utilizing various machine learning algorithms. The algorithms are Naïve Bayes, Maximum Entropy (MaxEnt), and Support Vector Machine (SVM). They remembered

emojis for the training data and utilized two classes for tweets' classification, positive and negative classes. In the wake of training data, they infer that emojis have a negative effect on data while applying MaxEnt and SVM algorithms on the data, however don't influence Naive Bayes. What is specific in their study is that, they explore the usage of unigrams, bigrams, combination of unigram and bigram and parts of speech. They conclude with the result the mix of unigrams and bigrams beats every other model, and parts of speech tags were not valuable at all.

3. Methodology

A. Sentiment analysis

With the usage of sentimental analysis, it can be learned whether the customers are satisfied with some new service or not. Twitter is mainly used for firms to get customer feedback. Simple articles are being written to identify whether people like or dislike something new. Firms are using that information to make a decision so that they can make some service better and improve the firm's sales. When sentiment analysis is applied on content, it means users are looking for the opinion in the text. Is the product review positive or negative? Are customers satisfied with the product or not? Are positive opinions greater than negative or not? All kinds of questions can be answered with Sentiment Analysis. By sentiment analysis, users can learn how customers' view the company's product or service. Shortly we can say sentiment analysis is being used for agree/disagree, like/dislike, for/against [10]. For example, the sentence 'I recommend this product to everyone.', the word 'recommend' indicates that the writer is happy, and the sentiment is positive.

In this paper, positive and negative words will be collected and used to train the machine to be able to classify the messages. For getting, storing, and classifying such data users will use Big data tools. Big data is data that exceeds the processing capacity of conventional database systems [11]. Big data means that there is a large number of data to collect. If users want to always get data from social sites faster, they should use big data. As data is more and more increased, it is becoming harder to control them, so Big data is the solution. Hadoop for years was the leading open source framework for Big data; recently Apache Spark is the leading and most popular framework. Hadoop and Spark almost perform the same tasks, but Spark is more preferable, especially when it comes to speed; because the way it processes data is faster.

B. Data and Findings

For the work and experiment, we used one document**.** The document contains different examples of messages with their outputs (classes) either positive or negative. The document is used to train and test the system because this computer program is going to be supervised learning, which is learning from example. They are using the known dataset for the training system called Stanford Twitter Sentiment Corpus (STS) [12]. Each tweet in this dataset has the following data: ID of the user, timestamp of the tweet, the username

of the user who posted the tweet, and the tweet itself. Next to each tweet, there is a class, either positive or negative. The document contains about 1 million samples of positive and negative tweets. In the following Figure, we show example of the dataset:

0	2329204987	Thu Jun 25 10:28:28 I	NO_QUERY	360cookie	Tried to get the mutant Fawkes tc
0	2329205009	Thu Jun 25 10:28:28 I	NO_QUERY	dandykim	Sick Spending my day laying in be
0	2329205038	Thu Jun 25 10:28:28 I	NO_QUERY	bigenya	Gmail is down?
0	2329205473	Thu Jun 25 10:28:30 I	NO_QUERY	LeeLHoke	rest in peace Farrah! So sad
0	2329205574	Thu Jun 25 10:28:30 I	NO_QUERY	davidlmulder	@Eric_Urbane Sounds like a rival
0	2329205794	Thu Jun 25 10:28:31 I	NO_QUERY	tpchandler	has to resit exams over summer..
4	1467822272	Mon Apr 06 22:22:45	NO_QUERY	ersle	I LOVE @Health4UandPets u guys
4	1467822273	Mon Apr 06 22:22:45	NO_QUERY	becca210	im meeting up with one of my bes
4	1467822283	Mon Apr 06 22:22:46	NO_QUERY	Wingman29	@DaRealSunisaKim Thanks for th
4	1467822287	Mon Apr 06 22:22:46	NO_QUERY	katarinka	Being sick can be really cheap wh

Figure 1. Samples of the Dataset

C. Process

First of all, we need to install Spark and include it in the Scala project. After that, we need to initialize a Spark Context, which is going to tell Spark how to access a cluster. The Spark Context takes a parameter, which is known as SparkConf or Spark Configuration. SparkConf allows the user to configure some common properties which will be passed to Spark Context, like application name, master URL. memory size, key value-pairs, and other properties.

```
val sparkConf = new SparkConf()
                .setMaster("local")
                .setAppName("TopHashtags")

val sc = new SparkContext (sparkConf) // An existing SparkContext.
```

Figure 2. Configuration

After configuration of the application, we started with the online collection of tweets. For online and real-time data, Spark streaming is required. Spark streaming receives live data from Twitter and divides them into batches, where the user can later apply actions and process the data. In the next figure, we show implementation of Spark Streaming.

```
val ssc = new StreamingContext (sc, Second (3))
```

Figure 3. Spark Streaming

User can get tweets from a specific secondary user, or all tweets that start with special word, or all tweets that contains special hashtag '#'. In our system, we collect all tweets containing special hashtag, and include that hashtag into the arguments of the system. Now, after all configurations we are able to collect data from Twitter. and save them to a file. In our system, we are saving the data to the text file. In the next figures, we show how to fetch data and how to save data into text files.

Journal of Natural Sciences and Engineering, Vol. 3, (2020)
DOI number: 10.14706/JONSAE2020114

```
val statuses = stream.map (status => status.getText())
```

Figure 4. Fetching Tweets

```
rdd.saveAtTextFile ('name of the file')
```

Figure 5. Save data in text format

D. Spark Machine Learning Library

The next and most important step is to classify each tweet to positive or negative class. Use Spark machine learning library, which contains different algorithms. Data, in order to be analyzed, it has to be converted to vectors. For that, use a well known and very useful tool called Hashing. Hashing is translating text data to numeric data. In Spark, most common and used hashing is HashingTF.it is important to say that, before analyzing the caught data from Twitter, it is a prerequisite to hash each data, as it is shown in the figure below.

```
val hashingTF = new HashingTF()
cal labelledTF = data.map {line =>
    val parts = line.split(',')
    val label = laebls(parts(0))
    val words : Seq[String] = parts(1).split("\\s+").map(_.toLowerCase
        )
    LabeledPoint(label, hashingTF.transform(words))
}
```

Figure 6. Hashing data

We used two algorithms for comparing the better one, Naive Bayes and Logistic Regression. Logistic Regression is a binary classification, which means it can classify data into one of two groups. While Naive Bayes can be used for multiple groups.First, we have used a 10 cross-validation. Cross-validation is splitting a dataset into more than one pan. It is used to ensure that every data has been used for training and testing data. Training data is always larger in size than testing data. If a user has 1000 samples of data, the user can take 800 for training and 200 for testing. Since he has used 10 cross-validation, it means 9 folds for training and 1 fold for testing.

Table 1. Cross validation example

1-fold	Training

2-fold	Training
3-fold	Training
4-fold	Training
5-fold	Ttraining
6-fold	Training
7-fold	Training
8-fold	Training
9-fold	Training
10-fold	**Testing**

Next, just move the testing data to another place in dataset, and another place in the table, like in table 2 where testing data is now 1-fold and it is at the top and beginning of the dataset. As we can understand testing data has to be moved each fold cross validation to one place and each data will be in testing and training part.

Table 2. Cross validation example 2

1-fold	**Testing**
2-fold	Training
3-fold	Training
4-fold	Training
5-fold	Ttraining
6-fold	Training
7-fold	Training
8-fold	Training
9-fold	Training
10-fold	Training

For each fold, it is important to calculate the accuracy; so, at the end you will determine its performance and if the classifier and data are good or not.Cross-validation and the accuracy are very important, they indicate to how well the learner will be able to make right and correct prediction for new data. For algorithms of learning, we used two machine learning algorithms as we mentioned before, Naïve Bayes and Logistic Regression. Results showed that Naive Bayes is better at prediction of the text. More details about the results will be described in the next section.

Journal of Natural Sciences and Engineering, Vol. 3, (2020)
DOI number: 10.14706/JONSAE2020114

4. Results

To train and test the system use Stanford Twitter Sentiment Corpus (STS) dataset which is available online. It contains more than one million samples. After the completion of testing on our data the results as well as accuracy of each k-fold is shown in the table below:

Table 3. 10-fold cross validation

k-fold	**Naive Bayes**	**Logistic Regression**
1-fold	77.3	68.8
2-fold	70.4	73.4
3-fold	75.7	74.3
4-fold	77.2	67.7
5-fold	76.4	64.6
6-fold	73.6	66.5
7-fold	69.8	75.3
8-fold	79.1	65.8
9-fold	77.3	67.2
10-fold	74.5	71.05

To calculate the accuracy of the classifier, true positive plus true negative over total number of testing data:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 7. Formula to Calculate the Accuracy

Code regarding our program:

```
val accuracy =
    1.0 * predictionAndLabel.filter(x => x._1 == x._2).count() / test.count()
```

Figure 8. Code to Calculate the Accuracy

'predictionAndLabel'- this is displaying the actual prediction result and the prediction of the system. Real example from our system is shown in the following figure, where it is shown the prediction of the system and real prediction of the data.

(0.0,0.0)
(0.0,0.0)
(1.0,0.0)
(0.0,0.0)
(0.0,0.0)

Figure 9. Prediction and Actual

Example: One sentence: ‘Good project, I liked it.’, result of classification using Naïve Bayes algorithm was: 1.0 means positive. while the result of LogisticRegression algorithm was: 0.0 which means negative sentiment. Another example: ‘I love it :)’, prediction of Naïve Bayes is 1.0 and the Logistic Regression is also 1.0 which is positive and correct.

The total accuracy of both algorithms, Naive Bayes and Logistic Regression, after cross-validation is shown in the following table.

Table 4. Accuracy

	Naive Bayes	Logistic Regression
Average Accuracy	75.13%	69.465%

From this table we can see that Naive Bayes average accuracy is somewhere around 75 percent. Logistic Regression accuracy is a bit lower than Naive Bayes and its accuracy is around 69 percent. There is some difference, not so big. That difference is around 6 percent. As a conclusion for those results **we take** the right to say that Naive Bayes algorithm provides great results. Logistic Regression with a this, bit lower percentage, can be considered as a great algorithm as well. After the users have finished the training of the system, use it for catching the data from Twitter and predict the data using both algorithms,Naïve Bayes or Logistic Regression. To get better results, we should use Naive Bayes rather than Logistic Regression. Finally, the best way is to save data in a text file, so the companies can easily keep track of the users' opinion about the company's products and about the company in general.

5. Discussion

In our paper, as you could see, we proved how text classification can be done in a fast and easy way by using Spark. Use Spark as Big data and for applying machine learning algorithms. Use two well-known machine learning algorithms, Naive Bayes and Logistic Regression. Using these algorithms we achieved a very high model's accuracy by applying to data sets that contained different types of sentences and emoticons. Also, we have shown how emoticons can help in improving the model's accuracy, if used correctly. Using more data in training and testing sets in our cross-validation method, we would achieve better results.

In this section of paper, an endeavor was made to compare the various methods and results of algorithms performance.Considering the research papers related to our research, which are already mentioned in Section 2, notice that in any case, the text should always be predicted using different methods and then decide which method is the best for achieving our goal. In the following table notice that, summarize different Supervised Machine Learning approaches for Twitter sentiment analysis.

Table 5. Summary of previous work

Paper	**Methods**	**Algorithms**	**Datasets**	**Results**
Pak and Paroubek [4]	Supervised Machine Learning	Multinomial Naive Bayes, Support Vector Machine (SVM), and Conditional Random Field (CRF)	Tweets collected using Twitter API	Multinomial Naive Bayes with bigrams accomplished a superior performance contrasted with unigrams and trigrams.
Go et al [9]	Supervised Machine Learning	Naive Bayes, Maximum Entropy (MaxEnt), and Suppor Vector Machine (SVM)	Tweets collected using Twitter API	The Maximum Entropy (MaxEnt) with both unigrams and bigrams accomplished a precision of 83% contrasted with the Naive Bayes with a precision of 82.7%.
Pang et al [2]	Supervised Machine Learning	Support Vector Machine (SVM), Naive Bayes, and MaxEnt	IMDb	The accuracy utilizing unigrams has 82.9% and accuracy utilizing the mix of unigrams and bigrams is 82.7% with Support Vector Machine (SVM). They proved that Support Vector Machine (SVM) is superior to Naive Bayes and Maximum Entropy (MaxEnt), where the accuracy utilizing unigrams has 81.0% with Naive Bayes and 80.4% with Maximum Entropy (MaxEnt), and the accuracy utilizing both unigrams and bigrams has 80.6% with Naive

				Bayes and 80.8% with Maximum Entropy (MaxEnt).

Some earlier research and studies utilized various groups of sentiment, similar to satisfaction, sadness, frustration, dread and shock. While, in our research, we classified the tweets into two groups, positive or negative, no third group. Most researches were about applying ML algorithms on tweets for sentiment analysis, without the use of Big data. While, we used Big data with the machine learning algorithms in our research.

From Table 5., see that Go et al got better accuracy using Naive Bayes algorithms. They did an additional procedure, which we neglected, and that is related to emoticons, they deleted any tweet that contains both positive and negative emoticons. This may happen if a tweet contains two subjects. Although we don't know the accuracy of the model in the research of Pak and Paroubel, we can surely say that they did a good research, because they followed the steps necessary to determine if the text is positive or negative. The steps followed included the removal of any URLs and usernames (user-names follow the "@" symbol) and removal of any characters that repeat more than twice turning a phrase such as OOMMMGGG to OOMMGG, which is applied by a regular expression.

6. Conclusion

In this paper it was shown how usage of Spark as Big data can help us classify text from tweets to positive and negative in a very simple yet very fast way.By using common algorithms Naïve Bayes and Logistic Regression we have achieved a very high by applying to large data sets that contained a various number of different emoticons and sentences. We determined that Naïve Bayes is much better than Logistic Regression by training and applying cross validation to our dataset, where its highest accuracy was around 79%. That is the most relevant result regarding the usage of Big Data. Also, in our paper we have demonstrated and shown how it is fast and easy to use and understand it, and how it is powerful with large data sets. For that reason, we can conclude that it is the best tool regarding Twitter sentiment analysis. But not only can sentimental analysis be used for Twitter, it can be used for any type of documentation or data. In the near future our plan is to have and use richer data sets for training, Spark Graphs for better data visualization and usage of real-time data rather than offline data. It can be achieved easy; just classification methods have to be applied and used right after getting each tweet from Twitter. We can see from the previous related works that are mentioned in the Chapter 2, sentiment analysis on Twitter data can be used in many different areas. From those papers, we can conclude that the main goal was to determine the products' quality, so we can say that the main goal is to make it easier for companies to check whether the item is good or not for the customers. Also, politicians and companies want to know what people write in real time about them, so they request monitoring tools to know the opinions, feelings and sentiments that their potential customers are publishing. This method can also be used in film production, since we can see that many Twitter users write their opinion about watched films, about the actors, and so on.

REFERENCES

[1] P. P. Chitturi, Apache Spark for Data Science Cookbook, Packt Publishing Ltd, 2016.

[2] B. Pang, L. Lee i S. Vaithyanathan, »Thumbs up? Sentiment Classification using Machine Learning,« 2002. [Mrežno]. Available: https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf.

[3] D. Kushal, S. Lawrence i D. M. Pennock, »Mining the Peanut Gallery: Opinion Extraction and,« 2003. [Mrežno]. Available: https://www.kushaldave.com/p451-dave.pdf.

[4] A. Pak i P. Paroubek, »Twitter as a Corpus for Sentiment Analysis and Opinion Mining,« 2010. [Mrežno]. Available: https://pdfs.semanticscholar.org/6b7f/c158541d5a7be2b2465f7d8a42afa97d7ae9.pdf?_ga=2.121841355.1543760336.1572899814-899645452.1571167125.

[5] Sanggeta, »Twitter Data Analysis Using FLUME & HIVE on Hadoop,« February 2016. [Mrežno]. Available: http://www.irdindia.in/journal_ijraet/pdf/vol4_iss2/27.pdf.

[6] J. P. Verma, B. Patel i A. Patel, »Big Data Analysis: Recommendation System with,« 2015. [Mrežno]. Available: https://www.researchgate.net/profile/Jaiprakash_Verma/publication/282686173_Big_Data_Analysis_Recommendation_System_with_Hadoop_Framework/links/57f4afb708ae280dd0b77681.pdf.

[7] K. R. Shrote i A. V. Deorankar, »Review based service recommendation for big data,« February 2016. [Mrežno]. Available: https://ieeexplore.ieee.org/document/7538334.

[8] F. Z. Ennaji, A. E. Fazziki, M. Sadgal i D. Benslimane, »Social intelligence framework: Extracting and analyzing opinions for social CRM,« November 2015. [Mrežno]. Available: https://ieeexplore.ieee.org/abstract/document/7507229.

[9] A. Go, R. Bhayani i L. Huang, »Twitter Sentiment Classification using Distant Supervision,« 2009. [Mrežno]. Available: https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

[10] l. Bing, Opinions, Sentiment, and Emotion in Text, Cambridge University Press, 2015.

[11] E. Dumbil, Big Data Now, O'Reilly, 2012.

[12] Kazanova, »Sentiment140 dataset with 1.6million twwets,« 2017. [Mrežno]. Available: https://www.kaggle.com/kazanova/sentiment140.

[13] C. t. W. projects, »Twitter,« 2007. [Mrežno]. Available: https://en.wikipedia.org/wiki/Twitter.

[14] A. Pak i P. Paroubek, »Twitter as a Corpus for Sentiment Analysis and Opinion Mining,« 2010. [Mrežno]. Available: https://pdfs.semanticscholar.org/6b7f/c158541d5a7be2b2465f7d8a42afa97d7ae9.pdf?_ga=2.121841355.1543760336.1572899814-899645452.1571167125.