International Journal on Data Science and Technology
2022; 8(1): 14-21
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20220801.13
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Data Mining and Revealing Hidden Sentiment in Tweets Using Spark

Ameen Abdullah Qaid Aqlan

Department of Computer Science, Kakatiya University, Warangal, India

Email address:
ameenaqlan218@gmail.com

To cite this article:
Ameen Abdullah Qaid Aqlan. Data Mining and Revealing Hidden Sentiment in Tweets Using Spark. *International Journal on Data Science and Technology.* Vol. 8, No. 1, 2022, pp. 14-21. doi: 10.11648/j.ijdst.20220801.13

Received: December 10, 2021; **Accepted**: January 22, 2022; **Published**: March 29, 2022

Abstract: Data science is important and scientific value in our lives because it is multi-fields, it's the science that uses scientific methods, processes, algorithms, and systems for the purpose of extracting knowledge and ideas from data whether this data is organized or not. Data science is called 21st century oil to highlight its importance and scientific value in our lives. We paid great attention in this research paper, where we achieved three main steps in the field of data analysis, collecting data from different sources in the Internet and then storing the data within the system, the second step cleaning the data in order to obtain structured data and then applying the algorithms that are responsible for classifying the data. In pursuit of development, we have collected more than 1600000 tweets about the educational process and the possibility of future online education. We give great attention to this field of Sentiment Analysis (SA) and the use of modern Spark technology, which has achieved great success since its emergence. We have succeeded in using Spark in getting a good result, handling data quickly and accurately, which encouraged us to test it on two algorithms of Machine Learning, Support Vector Machine (SVM) & Maximum Entropy (Max Ent).

Keywords: Sentiment Analysis, Maximum Entropy, SVM, Machine Learning (ML), Spark

1. Introduction

Big data represents an important stage in the development of information and communication systems, and in its simplified concept it reflects a vast amount of complex data that exceeds the ability of traditional software and computer mechanisms to store, process and distribute, resulting in the development of sophisticated alternative solutions that enable control and control of their flow.

Big Data technology has the ability to analysis internet site data, sensors and social network data, as the analysis of this data allows for links between a set of independent data to detect many aspects, such as forecasting the commercial trends of companies and combating crime in the security field and others [1]. These predictions also provide decision makers with innovative tools for better understanding of the circumstances and thus making the right decisions to achieve the desired goals. There are many large data sources, including those generated by the management of a program, whether a government or non-governmental program, such as electronic medical records, hospital visits, insurance records, bank records, food banks, etc. Commercial or transaction-related sources are another source, such as data arising from transactions between two entities, for example credit card transactions and transactions conducted via the Internet, including mobile devices. There are also sources based on sensor networks and tracking devices, for example, satellite imagery, road sensors, climate sensors and data tracking from mobile phones, GPS and others that can be a big data source. Another type of source is related to user behavior, such as internet searches for a product, service, or other type of information, and the times a page is viewed on the Internet. Finally, sources of opinion-related data such as comments on social media such as Facebook, Twitter and others. There are many tools and techniques used to analysis big data such as: Hadoop, Map Reduce, (Spark) is one of the most famous of these tools, it is an open source software program or platform written in Java language used to store and process big data in a distributed form, i.e. the storage of this big data is on several devices and then the processing is distributed to speed up these devices as a result of processing and return or call as a single package. The tools that deal with big data consist of three main parts: Data Mining, Data Analysis and, finally, dashboard tools.

1.1. Exploring and Defining Sentiment

Our research is great importance in light of the growing demand for data analysis and the utilization of the huge amount of it on social media sites and other sites, economically big countries like China and America are adopting huge investment stake in data wealth so that they can paint a clear picture of their future economies [2]. We will work to identify positive, negative and neutral sentiment and present some of these examples in Table 1. Many of developers gives a priority to studying positive and negative sentiment only without paying attention and considering a neutral sentiment. Here we will also study the neutral feeling and find out the correct number for it from the average tweets taken. Many tweets are useless and do not contain any sentiment, such these tweets are classified as neutral. That's why we have to deal with all tweets without exception to know what decision we have to decide [3]. These decisions will be at the heart and core of the management process and the success of the institution or government sector depends to a large extent on the ability and efficiency of the management leadership to make appropriate management decisions. The decision-making process begins with the collection, processing and debriefing of the decision-making process, on which many large companies and government sectors are beginning to rely on a complex and large data analysis policy that needs specialized data management and analytics software, which cannot be handled using only one tool or work on traditional data processing applications, it is known that data collection and information help to accurately characterize the problem and analysis it to reach accurate results, so it was necessary to adopt an administrative system that includes Analysis of big and very massive data. The government sector and large companies use the big data analysis system to improve internal processes, such as risk management and logistics service. It is also used to improve existing products and services, develop new services and products, utilize information and deliver appropriate customer presentations in a timely manner.

1.2. Query

We have a consistent assumption that users are interested in analysis of sentiment about the product and not the product, for example, when user ABC writes we do not evaluate the user's personality ABC but normalize the sentiment with ABC.

1.3. Characteristics of Tweets

Social media specially Twitter are a favorite of a wide range of users and have unique features that distinguish our research from the rest of the previous research. We explain some of the details of this comparison between our research and previous research.

Twitter is limited to 140 characters and this may be enough to convey an idea or express an opinion about something. The paragraph on Twitter is totally different from previous research that was interested in classifying reviews about the movies, previous works are also limited to movie reviews and Twitter reviews about different types of fields.

The process of collecting data and reviews from Twitter is simple and very easy to collect millions of them within minutes and this is the opposite of previous research that you do not exceed thousands of comments. we collected more than 1600000 tweets from twitter for analysis and give a good result. Most of the data collected is unstructured, there is some of users who writes his comments in the Slang or contains errors in the grammar,

Data preparation

Step1: downloads training data 1600000 tweets

<class 'pandas. Core. from. datafram' >

Count value, counts ()

String.value_counts ()

Step2: pre_clean Len (t)

Type: df. sentiment type

Description: sentiment class- 0: negative, 1: positive, 2: neutral

Step3: HTML Decoding

Step4: data filtering

Remove @

Remove space

Remove hashtag / numbers

Reduce URL Link

Step5: data cleaning function

Tweet cleaner (text):

Soup = Beautifulsoup text,

Souped = soup.get_taxt ()

Step6: testing

Testing = df. text [:100]

Cleaning and parsing the tweets.

nums = [0, 400000, 800000, 1200000, 1600000]

Step7: saving cleaned data

End.

Figure 1. *Algorithm 1.*

Table 1. *Example Tweets.*

Sentiment	Query	Tweet
Positive	learning	Selverdi: Distance learning will simplify several problems.
Negative	Everyone	Reysher: Not everyone can do that?
Neutral Neutral Negative	Subject Link website learning	Mairy: I didn't take my decide about this subject, link website classify as neutral Rami: distance learning will face problems.

2. Approach

Machine learning and data mining often use the same methods and overlap significantly, but while machine learning focuses on forecasting, based on "known characteristics" that have been utilized from training data, data extraction focuses on the discovery of properties (previously unknown) in the data (this is the analysis step for discovering knowledge in databases) [4]. Data discovery and exploration uses many machine learning methods, but with different objectives, machine learning also uses data mining methods as "unsupervised learning" or as a pre-treatment step to improve learner accuracy [5]. on the other hand, machine learning is enjoyed by a great deal of research centre where it is classified as one of the fastest and growing fields in the field of computer and data analysis, machine learning is one of the most important applications in the field of data analysis where it has become constantly supported and adapted to the technology of information [6]. We suggested using two different approaches of machine learning those are: Maximum Entropy, Support Vector Machines. We also used in feature extraction unigrams and bigrams, we use a framework that deals with classifiers and feature extractor as two distinct components. this works we built as a framework allows us to contribute and work better with different combinations of classifiers and feature extractors.

Emoji

Since the symbols are part of the tweets and comments write by users and express their conviction and opinions, we will pay great attention to detecting these emoticons and determining their emotion whether they are positive, negative or neutral, in the following table 2 we will display some of symbols most commonly used that are considered the opinions of their writers in the social media and other sites.

Table 2. *List of symbols.*

Emoticons mapped to positive	Emoticons mapped to neutral	Emoticons mapped to negative
:-)	-_-	:-(
:)	>:o	:(
:D	—,—	>:[
:o)	¯\^/¯	:<
:]		:-(

3. Machine Learning Approach

We have done a test for two different machine learning classifiers: keyword-based, Support Vector Machines, Maximum Entropy.

3.1. Support Vector Machines

SVM was officially disclosed and presented to the public in bose, guyon, 1992 through the fifth program of computer science, Themed the workshop was talking about data development and computational learning. we chose the SVM approach because it is the most popular and use as we can classify it among the most important techniques in data analysis and this is due to its success in the field of sentiment analysis. SVM is a discriminating approach that works in a different way, where it is defined by a super-separate hyperplane, we can say in another word, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples [7]. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either Sid. in the following Figure will illustrates main hyperplanes using with support vector machine, genetic algorithms (GA) classifiers and perceptron's on the two parts, and two class data [8]. Points within circles representing Support Vector, While the hyperplanes appear in different colors on the other side according to the pure lines in the figure.

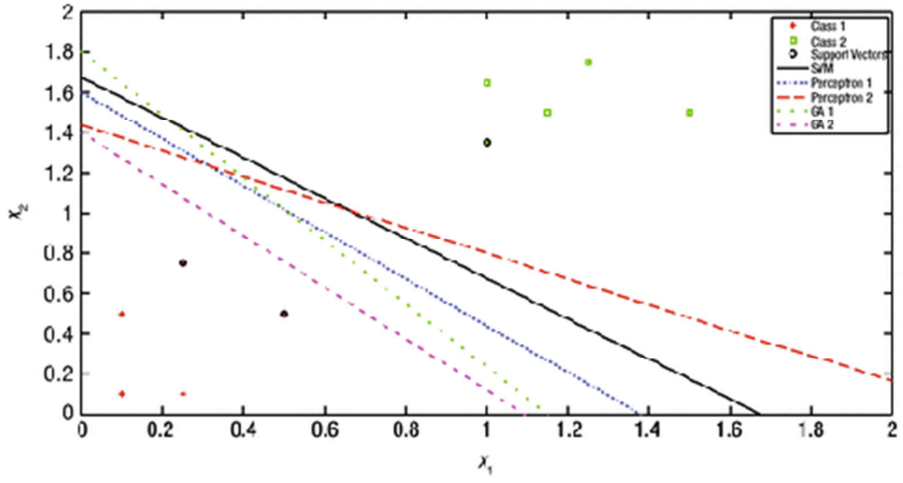

Figure 2. *Two-dimensional plot for SVM, GA and perceptron.*

3.2. Maximum Entropy

Max entropy principle expounded in 1957 by E. T. Jaynes in two articles, where he implicitly emphasized in his papers the existence of important approaches between information theory and statistical mechanics [9]. Classifier of maximum entropy is a very suitable in Text classification and giving solutions to problems affecting sentiment analysis, maximum entropy is probability classifier and more accurate in giving results and belongs to the exponential models' section. maximum entropy performs different tasks unlike the performance of Naive Bayes, which we have already discussed in detail in our previous work [10]. What distinguishes this work from the undo is that it does not builds any hypothesis that the feature is independent of each other [11]. We can use maximum entropy to solve variety of problems in text classification such as sentiment analysis, detection of language, and more. Max Entropy Principle is an important evidentiary principle, interested in distributing and identifying probability classes that enhances and enables the system to be categorized on the basis of appropriate and available data [12]. the Max Entropy is used in the range of two value 0 and 1 as we'll show that in the following figure 3.

Binary Entropy:

$$H(P) = -P \log P - (1-p) \log (1-p)$$

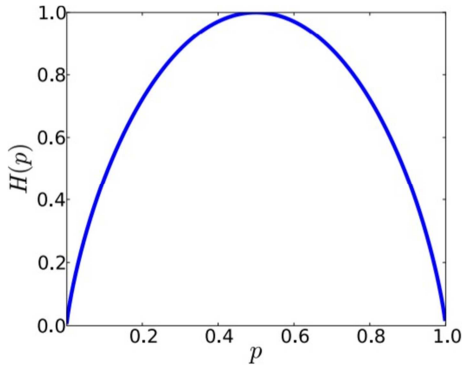

Figure 3. *Maximum Entropy Sigmoid.*

Framework of Max Entropy for Text Classification there are many features of maximum entropy principle, it can find the required possibilities through a special principle based on the creating as few assumptions, unlike other restrictions imposed, these constraints are created through a Consistent relationship between binary features and results [13]. The job of maximum entropy in Text Classification is assign a class called (g) from each word w, this principle derives its classification from documents (h) from the total training data (H). we Can compute the distributed x(g|h) as the following.

$$x\ (g|h) = 1\ y(h)\ \exp\ (\sum i\ \alpha i\ f\ i\ (h,g)) \qquad (1)$$

the following.

Where y(h) in equation (1) is an important normalization function which compute as the following.

$$y\ (h) = \sum_c\ e\ x\ p\ \ (\sum i\ \alpha_{\ i}\ f_{\ i}\ (h,g)) \qquad (2)$$

In equation 2 it shows us the parameter $\alpha_{\ i}$ with value, this value should give by estimation. We can estimate this in several manner and important studies in this area, we'll mention for example some important algorithms that we can use in our research, Broyden–Fletcher– Goldfarb–Shanno (BFGS), Improved Iterative Scaling (IIS) [14].

All of these algorithms represent a great success in this field, but we chose the first theory to use in our research as it is appropriate and more accurate. In equation 2, f_i (h,g), Represents advantage of binary value Represents advantage type equation here of binary value which support the Predict about the result (outcome).

$$Z_{wg}\ (h,g) = \begin{cases} 0 \\ \frac{N(h,w)}{N(h)} \end{cases} \qquad (3)$$

Where N(h,w) in equation 3 is the number of times word w occurs in document h, and N(h) is the number of words in h.

4. Hierarchy of Overall System in SA

Within a few years, the impact of big data has reached industries and activities ranging from marketing and advertising, to intelligence gathering and law enforcement, resulting in a lot of excitement and skepticism. Day after day, policy making looks like the next frontier for big data. Is this phenomenon, which expert Andreas Feigend called a 'new oil' to be refined [15]. With the huge explosion of data generated from small portions of 'deaf' data — such as numbers or facts — described by Alex 'Sandy' Pentland, a professor at The Massachusetts Institute of Technology, USA, as "digital breadcrumbs." [16]. A second type of big data includes videos, documents, blogs and other social media content. Most of this data is 'non- structural'; They differ from 'breadcrumbs' statements in that they are subject to the opinions of their authors, and may paint a deceptive picture because it is subjective. A third type of large- scale data is collected remotely by digital sensors, reflecting the actions of humans. These devices may be 'smart meters' installed in homes to record electricity consumption, or satellite images that can capture physical information, such as vegetation, as an indicator of deforestation [17]. Analysis technology Evolving remarkably fast, Developers are working on Employ machine learning algorithms to build models for relationship discovery and forecasting. In the Data Science Project, model selection and development are a complex process that requires expertise in areas such as computer science and statistics. Building a good prediction model means that it is provided with good data and the accuracy of the information it provides is highly accurate so that it uses statistical and mathematical methods and methods and comprehensive tests to ensure that the results of the model are logical and useful. All this is based on the questions asked in the initial stages of understanding the business and its needs.

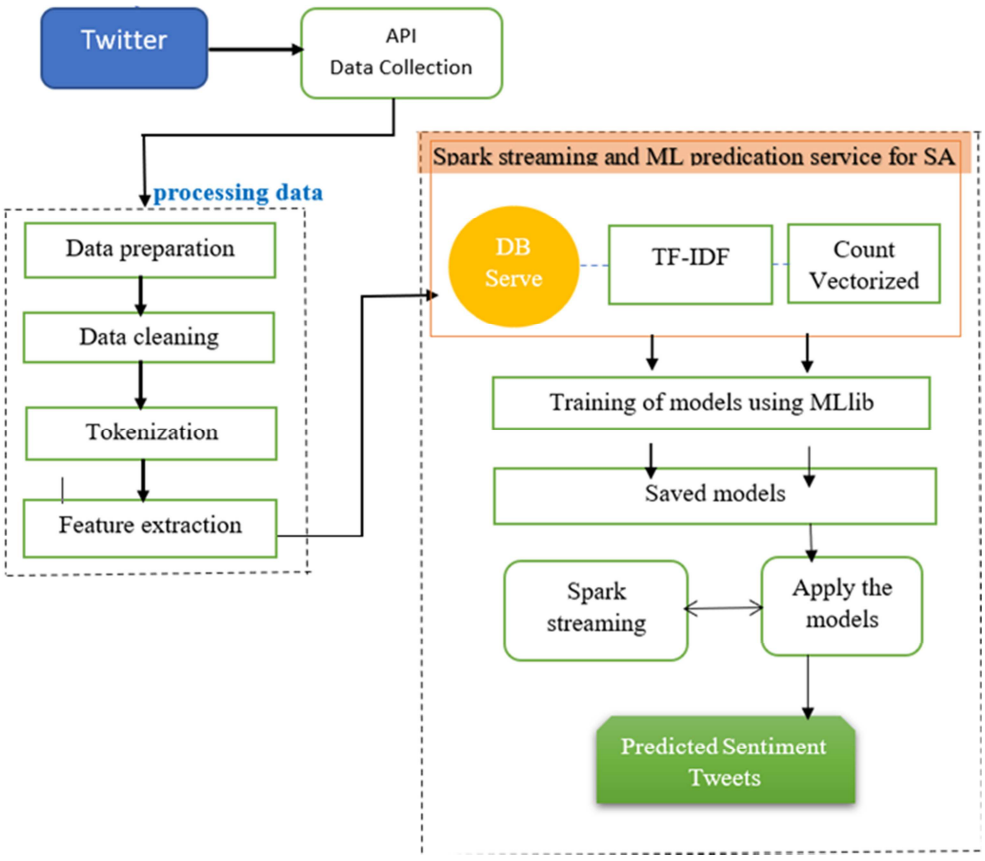

Figure 4. *Hierarchy of Overall System in SA.*

4.1. Data Collection

Twitter has tremendous data as it is a meeting place for all segments of society around the world making it the first platform for digital data. more importantly, Twitter allows developers to access user data via an application called API. after registering with the API app and accepting you as a developer allowing you to access the data, you are granted permission by OAuth Protocol to import the required data in accordance with a specific protocol and without sharing any information or privacy related to them or users. we have collected more than 1.6 lakh of tweets about students of Stanford and oxford university to know their opinion about the future of education as some universities move to adopt online study. all of these tweets collected are composite and unstructured, so we will use several tasks that enable us to obtain clean data.

4.2. Processing Data

First step is preparation of data set and save it on the system at a point called the information bank [18]. The second step is clean some symbols and text that has no value in the field of emotion analysis, such as HTML decoding, @mention, URL links, hashtag / numbers, space, UTF-8.

It seems like HTML encoding has not been converted to text, and ended up in text field such as &quot', '&amp','etc [19]. Decoding HTML to general and normal text will be my first step of data preparation and cleaning. we are going to use Beautiful Soup for this.

Second step of the data preparation is extract @mention and dealing with it in the right approach. Even though the @mention hold a specific information (which maybe another user that mentioned), most of this information doesn't give value to sentiment analysis model.

There are different types of encoding, UTF-8 is one of the most famous encodings in the devices used, the purpose is to encode the text [20]. There is other encoding that we can use it according to our programming. here we will decode UTF-8 and replace it in BOM (Byte Order Mark) so that we can identifying and processing it.

Sometimes we encounter a lot of space between words and this may create a gap when processing data, so we made a decision to control these spaces and delete them because they have no value and delete them does not affect the meaning.

The third step of the data preparation and cleaning is dealing with URL links, the same way we handled with @mention, although it carries information, we classify it as irrelevant in the field of sentiment analysis.

At the same time when we preparation and cleaning data we should focus on texts that contain the hashtag # before ignoring it, because most of the time it contains important and useful information and it is wrong to get rid of the entire text with the hashtag tag [21]. due to this reason we decided to deal with the text and numbers and delete the hashtag # only. We will apply this to all texts and numbers.

Tokenization will perform some tasks such as splitting the tokens and other tasks. it will be dealt also with next stage when creating matrix with either Tf-idf vectorizer or count vectorizer.

Feature extraction: After completing the stage of preliminary processing and keep of the data ready, we move to the second stage of implementation and training by Max Entropy Classifier on TF-IDF [22] weighted word frequency features, It also explores distinctive words in all files without interest in the subject of feelings, words such as like, dislike, good, bad, distance, interested and other, same this words will be taken as features of the word. When the term (word) is repeated more, the weight of the term is increases. This is the work of

$$\text{tf- idf tf}(t,d) = \log(F(t,d)), \quad (4)$$

where F(t,d) Indicate to the repeated of the terms (word)

$$\text{idf}(t, D) = \frac{\log(\frac{N}{N})}{t \in d} \quad (5)$$

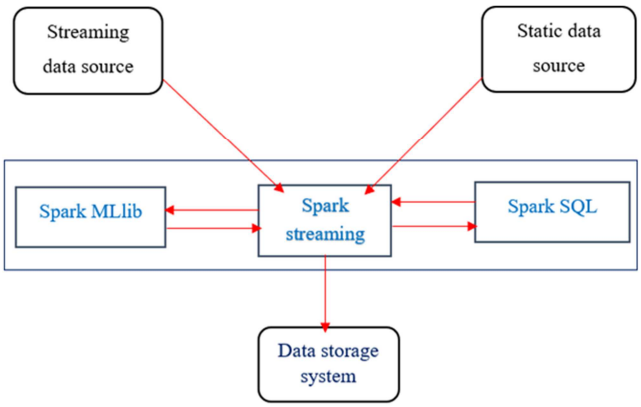

Figure 5. *Procedure of Spark Streaming.*

4.3. Spark Streaming and ML Using in SA

Spark set a world record by completing a benchmark test involving sorting 100 terabytes of data in 23 minutes - the previous world record of 71 minutes being held by Hadoop, (Bernard Marr) [23]. Python is a mission in data science, thanks to number of reasons, including the abundance of modelling in data science. Python support different data set, and it's also easy to use and open source. But what if we deal with huge and inappropriate data on python? Here we can use Apache of spark because it can handle unlimited data and execute with high efficiency. Recent studies by most developers and technicians in this field indicate that Spark is one of the most advanced technologies of our modern era. use of spark in the present time along with machine learning technology has positive implications, as studies have shown that Spark is very suitable for machine learning technology. Spark uses computing (computational) due to its ability to calculate analytical calculations and this benefits us a lot in processing and analyzing data with less time and high accuracy. the approach we used it in sentiment analysis by Spark show as Positive results in many respects, the data has been processed is estimated at 1,600,000 tweets, where we got Final test, accuracy is 81.22%. This result is good compared to previous work that has not recorded such a number, where we can increase the number of data to be analyzed and this does not affect accuracy.

4.3.1. Count Vectorized

In order to take advantage of the textual data, we should build suitable approach helps us to refine and encode the words correctly. CountVectorizer takes care of in this process which is called vectorization, All the targeted words in the text need coding and numbers for use as inputs in machine learning algorithms [24].

4.3.2. Training of Models Using ML

Step1: reading the data from file

Step2: collect the train set data from pc

Step3: collect all the tokenizer & share on tokenizer variable

Step4: letter extract random & select featured using HashingTF function & save in HashTF variable

Step5: remove all spaces terms & score in label-shingIndexer

Step6: perfume's pipeline function by tokenizer, hashTF, Label, StringIndexer pipeline (n., n., n)

Sterp7: later less than data frame on the machine learning alg using pipelinefit () & pipelinefit. Random () functions

Step8: later case machine learning alg. Maximum Entropy ()

Step9: later case Binary classification

Sterp10: main Accuracy.

End.

Figure 6. *Algorithm 2.*

We will create the most important step in this part of sentiment analysis series, focusing on the parallel work of both the ML and Spark in building the model. This model can predict Sentiment and classify them into three polarities, positive, negative and neutral. We're going to build this model on important and more complex functions in this part of the series. Moving to the world of social media and big data requires us to have more modern applications such as Spark and Machine learning, distributing Spark with machine learning give us a library of shared algorithms capable of working well. This distribution or approach attracts a lot of developers due to its low cost rather than other high-cost distributions and this helps a lot to grow and develop this field faster.

4.3.3. Predicted Sentiment Tweets (Result)

After all these important details we extracted the percentage of both positive, negative and neutral opinions, this percentage shows us an unfatused number who refuses to teach remotely, while the second category has taken the neutrality towards this, and a few of them support the online study process. In the following figure 7 will show the polarity of different types reviews.

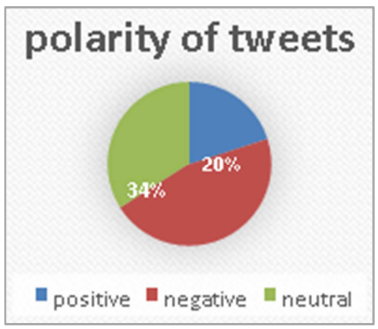

Figure 7. *Number of Positive, Negative and Neutral Tweets.*

SVM & Max Ent, both of them give better results in sentiment analysis, but with the use of mode rn Spark we have obtained very good results compared to previous works using python and others.

The difference was simple between SVM & Max Ent algorithm, in SVM we got Accuracy Sco re: 81.22 with ROC-AUC: 88.62, & Max Ent 80.00 with 88.62.

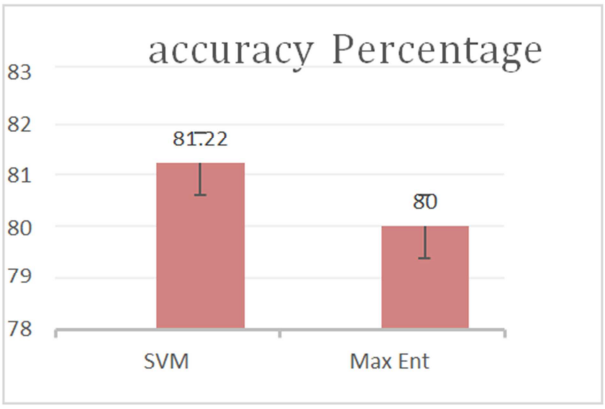

Figure 8. *Accuracy percentage of SVM & Max Ent.*

Table 3. *Average Result of SVM& Max Ent.*

ML Algorithm	Accuracy	ROC-AUC
SVM	81.22	88.62
Max Ent	80.00	88.62

5. Conclusion

We have collected more than 1,600,000 tweets from to process them and improve the field of sentiment analysis according to a new approach. Our modern approach is to take advantage of Spark technology that is characterized in Lightning-fast processing speed and others, Spark is a big data analysis tool that enables data scientists to build more accurate and faster models, spark is an open source. We have used two machine learning Maximum entropy and support vector machine, where we got a clear improvement in terms of speed and results. in SVM we got Ac- curacy Score: 81.22 with ROC-AUC: 88.62, & Max Ent 80.00 with 88.62. Good results prove to us how important use modern technology such as Spark with Machine Learning in development field of sentiment analysis.

References

[1] Kumar, Akshi, et al. "Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network." *IEEE Access* 7 (2019): 23319-23328.

[2] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams engineering journal* 5.4 (2014): 1093-1113.

[3] El Alaoui, Imane, et al. "A novel adaptable approach for sentiment analysis on big social data." *Journal of Big Data* 5.1 (2018): 12.

[4] Hemalatha, I., GP Saradhi Varma, and A. Govardhan. "Sentiment analysis tool using machine learning algorithms." *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* 2.2 (2013): 105-109.

[5] Kawade, Dipak R., and Kavita S. Oza. "Sentiment analysis: machine learning approach." *Int. J. Eng. Technol.(IJET)* 9.3 (2017).

[6] Osisanwo, F. Y., et al. "Supervised machine learning algorithms: classification and comparison." *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017): 128-138.

[7] Yang, Yong, Chun Xu, and Ge Ren. "Sentiment analysis of text using SVM." *Electrical, Information Engineering and Mechatronics 2011*. Springer, London, 2012. 1133-1139.

[8] Awad, Mariette, and Rahul Khanna. *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Springer Nature, 2015. https://en.wikipedia.org/wiki/Principle_of_maximum_entropy

[9] De Martino, Andrea, and Daniele De Martino. "An introduction to the maximum entropy approach and its application to inference problems in biology." *Heliyon* 4.4 (2018): e00596.

[10] Liu, Xiao, et al. "Pricing Interval European Option with the Principle of Maximum Entropy." *Entropy* 21.8 (2019): 788.

[11] https://www.tsia.com/blog/the-new-data-refineries-transforming-big-data-into-decisions

[12] https://www.unglobalpulse.org/2011/11/social-impact-through-satellite-remote-sensing-visualizing-acute-and-chronic-crises-beyond-the-visible-spectrum

[13] Ameen Aqlan, and other, "A Study of Sentiment Analysis: Concepts, Techniques, and Challenges" doi.org/10.1007/978-981-13-6459-4_16

[14] Jaynes, Edwin T. "Information theory and statistical mechanics." *Physical review* 106.4 (1957): 620.

[15] Nigam, Kamal, John Lafferty, and Andrew McCallum. "Using maximum entropy for text classification." *IJCAI-99 workshop on machine learning for information filtering*. Vol. 1. No. 1. 1999.

[16] Berger, Adam, Stephen A. Della Pietra, and Vincent J. Della Pietra. "A maximum entropy approach to natural language processing." *Computational linguistics* 22.1 (1996): 39-71.

[17] Bao, Yanwei, et al. "The role of pre-processing in twitter sentiment analysis." *International conference on intelligent computing*. Springer, Cham, 2014.

[18] Teufl, Peter, and Stefan Kraxberger. "Extracting semantic knowledge from twitter." *International Conference on Electronic Participation*. Springer, Berlin, Heidelberg, 2011.

[19] Effrosynidis, Dimitrios, Symeon Symeonidis, and Avi Arampatzis. "A comparison of pre-processing techniques for twitter sentiment analysis." *International Conference on Theory and Practice of Digital Libraries*. Springer, Cham, 2017.

[20] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre-processing in sentiment analysis." *Procedia Computer Science* 17 (2013): 26-32.

[21] Htet, Hein, Soe Soe Khaing, and Yi Yi Myint. "Tweets sentiment analysis for healthcare on big data processing and IoT architecture using maximum entropy classifier." *International Conference on Big Data Analysis and Deep Learning Applications*. Springer, Singapore, 2018. Bernard marr, big data in practice. Springer Journal.

[22] Jodha, Rajshree, et al. "Text Classification using KNN with different Features Selection Methods." *Text Classification using KNN with different Features Selection Methods* 8.1 (2018): 8-8.

[23] Guller, Mohammed. *Big data analytics with Spark: A practitioner's guide to using Spark for large scale data analysis*. Apress, 2015.

[24] Shakhovska, Natalya. *Advances in Intelligent Systems and Computing*. Springer International Pu, 2017.