T.C.

ALTINBAS UNIVERSITY

INFORMATION TECHNOLOGY

SENTIMENT ANALYSIS IN DATA OF TWITTER USING MACHINE LEARNING ALGORITHMS

MUSTAFA AHMED MAHMOOD

Master Thesis

Supervisor:

Asst. Prof. Dr. Sefer KURNAZ

Istanbul, (2019)

[SENTIMENT ANALYSIS IN DATA OF TWITTER USING MACHINE LEARNING ALGORITHMS]

by

[Mustafa Ahmed Mahmood]

Information Technology

Submitted to the Graduate School of Science and Engineering in partial fulfillment of the requirements for the degree of Master of Information technology

Master of Science

ALTINBAŞ UNIVERSITY

[2019]

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

____

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Prof. Dr. Osman N.UCAN	School of Engineering and Natural Science, Altinbaş University	____________________
Asst. Prof. Dr. Sefer KURNAZ	School of Engineering and Natural Science, Altinbaş University	____________________
Asst. Prof. Dr. Zeynep ALTAN	Beykent University	____________________

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

____

Asst. Prof. Dr. Oguz Ata

Head of Department

____

Assoc. Prof. Dr. Oguz Bayat

Director

Approval Date of Graduate School of
Science and Engineering: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

[Mustafa Ahmed Mahmood]

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Sefer KURNAZ for his supervision, and I thank him very much for helping me to complete this thesis on time. Thanks to all the doctors in information technology department.

ÖZET

[HESSDS ANEILIZLERININ TWITTER VERILERINDE KEULLEHILER MAKENSI ALGORITME LERENIN ÒĞRENIMI]

[Mahmood, Mustafa Ahmed]

[Yüksek Lisans], [Bilişim Teknolojileri], Altınbaş Üniversitesi,

Danışman: Dr. Sefer KURNAZ

Tarih: [Mart, 2019]

Sayfa Sayısı: 58

Sosyal medya kullanıcıları tarafından yazılar, tweetler, resimler ve videolar gibi her saniye için büyük miktarda veri üretilir. Bu büyük verilerden değerli bilgi almak, metin madenciliği alanında önemli, zorlu ve ilginç bir konudur. Twitter verileri, toplum gündemini, eğilimlerini, kullanıcı davranışlarını ve duygularını keşfetmek için metin madenciliği teknikleriyle analiz edilir. Tweet'lerden gelen duyguları belirlemek için bir metin analizi yöntemi önerdik. Verileri anlamlı bir ortama koymak için doğal dil işleme teknikleri uygulanmaktadır. Bundan sonra sınıflandırma modeli, işlenen veriler üzerinde veri madenciliği yöntemleri ile eğitilir. Twitters akış verilerini kullanarak sınıflandırma etiketini, olumlu, olumsuz ve tarafsız duygular gibi, insanların görüşleri olarak gerçekleştirir. Twitter API kullanarak imdb'yi seçiyoruz ve bu markalar hakkında hashtagleri olan tweetleri topluyoruz.

Anahtar Kelimeler: Metin madenciliği, duyarlılık analizi, tweet analizi.

ABSTRACT

[SENTIMENT ANALYSIS IN DATA OF TWITTER USING MACHINE LEARNING ALGORITHMS]

[Mahmood, Mustafa Ahmed]

[M.S],[Information Technologies], Altınbaş University

Supervisor: Prof. Dr. Sefer KURNAZ

Date: [Mars 2019]

Pages:58

Massive amounts of data are generated by social media users for each second, such as posts, tweets, images, and videos. Getting valuable information from this big data is a significant, challenging and interesting issue in the text mining area. Twitter data are analyzed with text mining techniques to discover society agenda, trends, user behaviors, and feelings. We proposed a text analysis method to determine sentiments from tweets. Natural language processing techniques are carried out to put the data into meaningful context. After that classification model is trained with data mining methods on the processed data. It carries out the classification label as people's opinion, such as positive, negative, and neutral sentiments, using Twitters streaming data. We select imdb and collect tweets with hashtags about these brands by using twitter API.

Keywords: Text mining, sentiment analysis, tweets analysis.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

API	: Application Programming Interface
ICT	: Information and Communications Technology
IR	: Information Retrieval
NLP	: Natural Language Processing
OAuth	: Open Authentication

1. INTRODUCTION

These days, the quantity of the statistics flowing the use of the net it extended very drastically. This issue has made innovations and discoveries within the area of facts and communications technology (ICT) to generate a brand-new environment able to handle a big record. To our gift time generation has reached a level, in which you discover humans are interconnected with every different and with the approach of social verbal exchange and are able to percentage their lives and their feelings via social networking. A few sparkling demanding situations came into view concerning architecture of facts storage with scalability characteristics and algorithms with efficient processing. Twitter is a critical tool for supplying facts for tens of millions of people, consequently it is appropriate to be implemented in data mining field. Much interest changed into directed towards the idea of community in the global of social networking. Studying twitter facts with these algorithms could be very difficult because it is a micro running a blog offering with asymmetrical functions. If you will observe me, I’m no longer obliged to comply with you. A whole lot of users in twitter have greater followers than they are privy to, which imply that twitter connections is less depending on in-individual touch. It's miles very useful to study the twitter environment to recognize how individuals use these recently evolved communication technologies for the motive of setting up social connections and maintaining the present ones. We begin with a few studies on how twitter geo-tagged tweets may be utilized for the motive of figuring out essential consumer behaviors and capabilities a as well as figuring out places of pursuits/landmarks. After that we recommend numerous similarity degree for detecting groups and introduce methods for reading class algorithms. Currently, the social media developed as a supply of real time, political and social data. additionally, it's miles a prime means of advertising and communique. customers are sharing their information at the social networks through using running a blog, statuses, sharing films and pox similarly to interacting with each different, consequently they may be able to generate communities and groups on social network. Analyzing and monitoring this fact may want to cause sizeable insights that in different aspects might be tough to apply media resources and conventional techniques. Social networking web sites like twitter, flicker and fb introduce a sparkling means for purchasing common updates and sharing facts among them. Additionally, those websites allow the sharing of greater information that would be of significance in content reading, e.g. place et carta.

One benefit for the social media over the traditional media resources, that its miles maintained and controlled through customers. Conventional media totally authorized the customers to have the statistics which changed into offered to them. Information drift was one sided that means it flows from media supply to customers. At the same time as in social community the consumer has the functionality of responding to the occasions and information surrounding them also offering their thoughts and sharing them. That led to growing multiway mode of spreading the facts wherein the user publishes records with further information consisting of videos, pictures and links. This ends in producing facts version this is generated with the aid of the user. The consumer's social graph and the consumer connections in terms of the social network have a big role within the system of reading the records version to get essential data from huge quantity of "user generated content material" that's generated on day by day basis. As, microblogging sites which include flicker, twitter and Facebook permit the user to share shortened multimedia and messages, these web sites become an immediate statistics supply where the users all over the global are capable of stay linked and feature numerous resources to achieve information [1].

In addition, twitter is verbal exchange platform globally from all around the global. Twitter has been developed to hold tempo with the fast verbal exchange among the arena. Its flexibility and speedy tracking have made changing messages and twitter among humans one of the maximum critical contacts. Data is accumulated from twitter thru an application programming interface [2].

One of the most significant twitter features which distinguish it from other social networks like Facebook is that the relation between the one being followed and the one who is following does not need to be two-ways. Following a user in twitter is the same as subscribing to blog. The user who is following get all twitter status updates of the user he follows. Twitter can be considered as a pretty good reaction to all what happens in the worlds since it is widely used in all daily life aspects. Among all the things that happening, most recent trends are of a great interest to the companies. Most recent trends could be put to analyzing and when reacted to identified. From marketing perspective, these most recent trends could be utilized for the purpose of responding with suitable activities, such as advertising products. Tweets analyzing could be of a significance for the companies for generating an advantage to rivals [3].

1.1 IMPETUS

In the current associate world, customers' ability sends messages whenever they want. Yet, social media cannot be considered as a typical mean for sharing and messaging private thoughts and things; it is utilized through public figures, journalists and politicians, universities and corporations who tend toward reaching more public and sharing their ideas and have interest in the views of individuals. The energetic increase of persons using the social media caused the development of such resources as a source of reflecting views and mood of individuals. Tracking the reactions of individuals in social media throughout crises drew a rising level of attention in the research society [4].

The public's sentiment analysis is strongly criticized in macroscale socio-economic phenomena such as the prediction of stock market rate of a company. This might be achieved via examining the total public sentiment concerning that company regarding time and utilizing the tools of economics to find the association between the stock markets rate of the company and the public sentiment. Also, companies might assess the response and performance of their product in the market, in which area this product is having undesirable and where it is having positive response in the market (because twitter offer downloading stream of geo-tagged tweets for locations [5].

Analysing and monitoring sentiments on microblogs, particularly Twitter, offers huge prospects for the private and public sectors. For example, in the private sector, it was noticed that the status of a capital stock, firm or product is largely impacted via the sentiment and rumours shared and published among individuals on social networks and microblogging platforms [6].

Researchers noticed that a rage number of publications left by individuals each month, could not be processed in a manual way via conducting public opinion polls. Such situation draw attention to the necessity for automated approaches regarding the intellectual analysis of text information, and that allow processing huge amount of data in short period as well as understanding the messages of users. The most complex and vital element regarding automated processing is understanding the messages of the users. Utilizing current approaches and technologies of big data, applying artificial intelligence is being of a great help to researchers for the purpose of automating the process of content analysis, collecting data, also for preparing, managing and visualizing visualize data. These improvements offer great advantages in monitoring social media in real time and conducting extensive researches [7].

1.2 QUESTION OF RESEARCH

The next main research question is concluded according to the observations which have been considered above.

RQ 1. How to Ways the methods of (ML) used for Rating feelings comparing in the information of social media?

Certain methods that are considered in Ch2 aren’t usually applied to function with text data. For the purpose of applying such models, data must be altered, text preprocessing must be achieved. This give rise to the second question in the presented study:

RQ 2. What sort of preprocessing approaches are presented to convert natural language text into appropriate format?

For the purpose of answering such questions, sets of experimentation should be performed with natural language approaches as well as machine learning methods.

1.3 SOCIAL NETWORKING SITES

Web based life may be considered as a gathering of web put together applications that work with respect to the ideological and innovative establishments of Web 2.0, and that permit the creation and trade of client produced content", as stated by Kaplan and Haenlein [8].

Recently, along with the front-runners of the Web like Twitter, LinkedIn also fb, exist innovative services for various groups of individuals: students' social network, the networks intended for certain groups of ethnic minorities or specialists, in addition to an unusual network for all the drinkers in the world. This ranges the extent to extremely distinct types of research from customer favorites to psychological features. Facebook kept the leader among social platforms in the first few months of 2015, while twitter has been in the top 10. Based on the same research made by Simon Kemp [9].

Globally, over two billion individuals are considered as active users of blogs and social networks. The total social media landscape is dominated via Facebook, maintaining 1.366 billion actives users in Jan 2015. In the meantime, chatting applications and instant messengers continue to develop, with Viber, WeChat, WhatsApp and Facebook Messenger indicating over one-hundred million active users each month over the past 2014. Today, chatting applications and services of instant messenger make up three of the top five global social platforms, and eight brands of instant messengers indicate over one-hundred million active users each month. In 2015, the number of active users in twitter each month has been 284, while in 2016 the number surpassed 320 million active users each month [10].

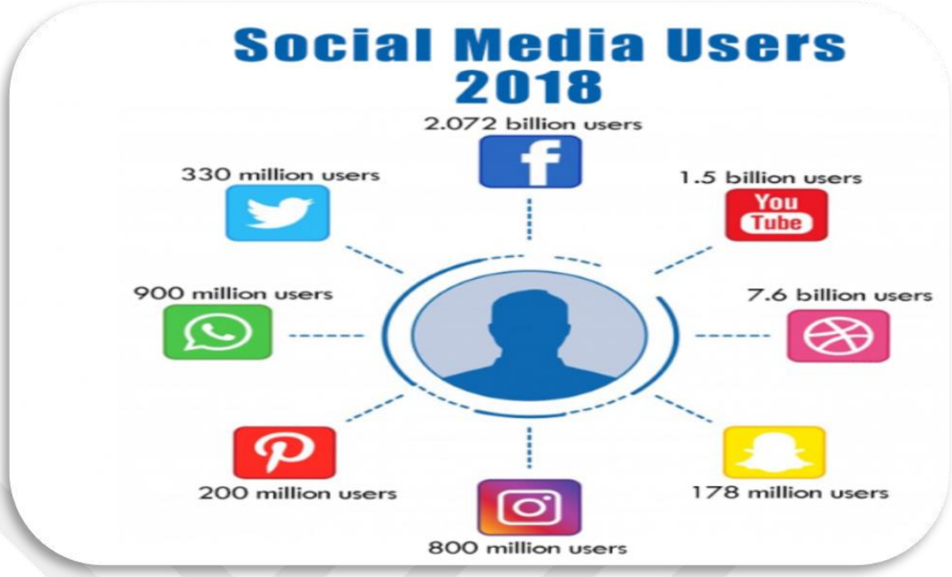


Figure 1.1: accounts user active in social media site in 2018 [11]

1.3.1 The Specifics Content of Twitter

Twitter is the most utilized micro-blogging site with “500 million” users and “340 million” tweets daily, that is why it is an important information source. Twitter differentiated from other social sites by their restricted in size messages. There is a maximum length of 140 characters per tweet which makes challenging to detect sentiment from such short and informal comment [10].

The tweet is a way of sharing interests among specific groups or publicly. These Tweets are posted through different sources such as twitter mobile applications, twitter site, and several third-party websites/applications (after authenticating process). Privacy features are controlled by the user, as users can decide whether to make their tweets private and only the people who have permission from the user are able to access these tweets, or to make the tweets publicly visible. Twitter users can follow other users that give them the ability to access tweets on their twitter page [11].

Twitter have several other features. One feature permits the user to reply to other users tweets by choosing to click the reply button on the tweet made by another user. It is a way were a user can respond to the tweet made by another user. Additionally, a ‘@’ symbol is added before the username a twitter user when another user wants to mention the first one in his tweet. Mentioning is a way where a user wants to refer to another one. Retweeting is another twitter common concept. Retweet means to share other user tweet to our followers, it plays a significant role in spreading information within twitter. Also, a Hashtag ‘#’ sign is added by the user before relevant keywords in their tweets. This is utilized for the purpose of categorizing the tweets to be shown easily in the twitter search. Popular twitter hashtags turn to be trending subjects in twitter [4].

1.4 SENTIMENT ANALYSIS

Sentiment analysis is defined as a field of natural language processing (NLP), it has a task to determine attitudes and views of writers in the text or their opinions regarding certain subjects. Sentiment defines the views and opinions which were stated through a person (the view holder) regarding an entity (target). Attitudes quite enduring, effectively colored opinions, preferences, and tendencies regarding individuals or objects (hating, loving, liking)"are considered to differ from sentiments "brief episodes of synchronized responses (angry, sad, joyful, fearful, ashamed, proud)" as a response to external impacts [15].

This differentiates sentiment analysis from the other issues including the analysis of emotions in which the general emotional state (impacted via a lot of external aspects) is of high importance, not the outlook regarding a target. The direction and degree of sentiment (how negative or positive it is) is referred to as its polarity. The easiest and most general polarity approach supposes 2 types, negative and positive. The extreme ends of continuous or discrete scale are constituted via these 2 types. Such description involves most voting approaches that are utilized in practice, including:

- positive, neutral, negative (such as: eBay),
- thumbs up/down (such as: Facebook, YouTube),
- star ratings (such as: Amazon, IMDb).

Frequently, polarity will be mapped to [-1; 1] interval, in the case when -1 is considered as the most possible negative polarity, and 1 considered as the most positive polarity. Some ambiguity is presented in relation to the center of the scale (0), that is generally defined as neutral. Yet, this could also refer to a balanced mixture of negative and positive content [16].

1.4.1 Types of Sentiment Analysis

In today's research, various viewpoints regarding automatic sentiment analysis are presented, which lead to different tasks. The granularity of analysis is considered as the most noticeable difference between them. Analysis of sentiments is implemented on numerous linguistic levels. At sentence level, the aim is classifying if the sentence of an individual has neutral negative or positive sentiment, while at document level, the main aim is classifying if a whole opinionated document has neutral, negative or positive sentiment [17].

At aspect (entity) level, the aim is classifying the sentiment of person’s phrases or sentences aimed at the direction of aspects or entities. The task of document-level analyzing of sentiments is predicting the total polarity stated in documents. Usually, the documents where this sort of analysis is achieved are the documents where the writer assesses just a single entity, like reviews related to movies, hotels or products. The prediction of document-level polarity could be developed as standard text classification issue. The issue could be dealt with using ML methods, including Naïve Bayesian classifier and maximum entropy classification. There are some statements implicated in the text classification method. The first one assumes that the entire text is related to single target, that is the product which might be defined as the topic of the review. The second one assumes that the author is the opinion holder. In a formal way, the task of document-level classification of sentiments is defined in the following way [18].

A group of documents D exists with the writers’ views, it concludes if a document d ϵ D express a negative or positive view (or sentiment) on a topic. If there is a document d that makes remarks on a certain object o, determines orientation oo regarding an opinion stated on o, in other words, determine the opinion orientation oo on feature f in quintuple (o; f; oo; h; t), where f = o and h, t, o is presumed to be unrelated or identified. The next hypothesis is made by recent studies regarding sentiment classification. Document d (for example, a product evaluation) express sentiments on an object o and opinions are from single opinion holder h. This statement applies for review made by customers regarding services and products. Nevertheless, it might not apply for a blog post or forum because in a post like that the writer might express feelings on various merchandises and compare these products utilizing superlative and comparative sentences. Usually, the prediction of the polarity of sentence is considered as a classification task [19].

1.5 MACHINE LEARNING TECHNIQUES

There are 2 central methods in sentiment analysis (unsupervised and supervised methods). Sentiment classification related to the data in twitter is achieved via the use of supervised ML methods such as Maximum-Entropy, SVM and Naive-Bayesian. The classifier's effectiveness classifier is built on which dataset is applied for which classification approaches. Regarding supervised machine learning methods for training the classification model, training dataset is applied which support classifying the test data [20].

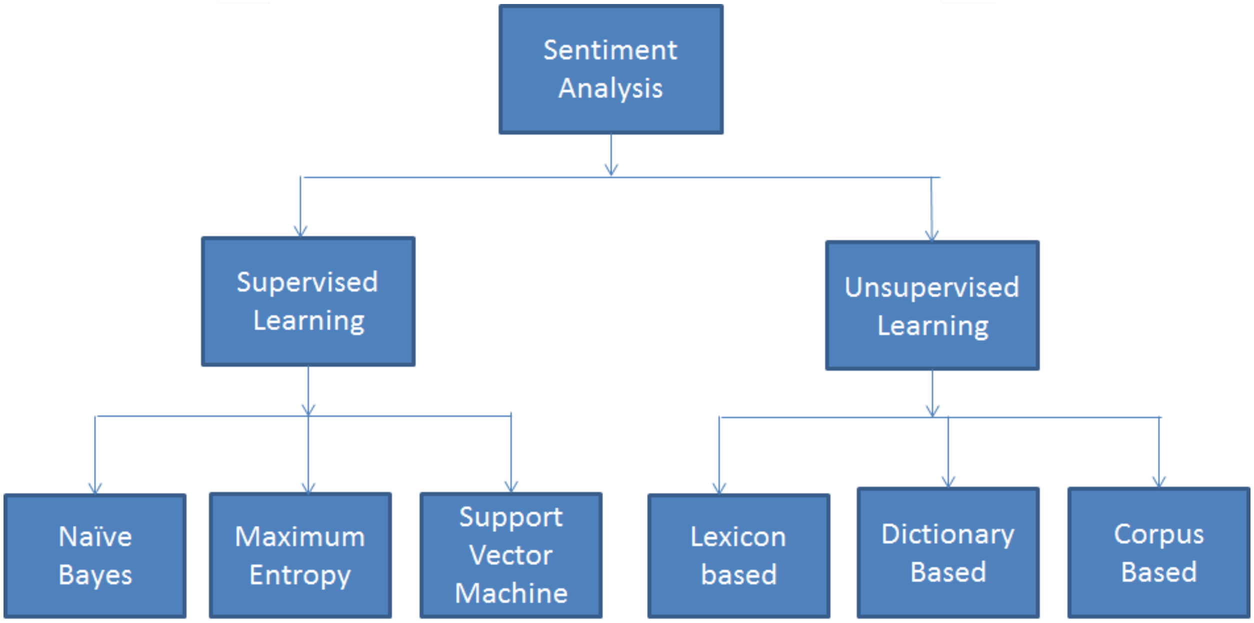


Figure1.2: Sentiment Analysis Algorithms

1.5.1 Naive Bayes

Naive Bayes classifier can be considered as a simple classification model and it functions properly on text classification, also it is probabilistic classifier that depends on using Bayesian theorem with solid assumptions of independence. This is the most straightforward type of Bayes Network, where all attributes are independent assuming the value of the class variable. Which is referred to as the conditional independence. It supposes that every one of the features is conditional unrelated to other features given the class. Naive Bayesian classifier can be defined as a method which is applied to a particular class of issues, namely the ones that are phrased as connecting an object with a discrete class. From numerical based approach group, Naive Bayesian has some benefits for example it has high-accuracy, fast and simple, in K. Ming Leung defines the Bayes rule [21].

$$\mathrm{P(Y|X)} = \frac{\mathrm{P(X|Y)\ P(X)}}{\mathrm{P(X)}} \tag{1.1}$$

In the Naive Bayesian classifier, X will speak to the in-position highlight and Y will speak to the class variable. Let X (x1; x2; : ; xm), where xi is speaking to the estimation of highlight I. Let (y1; y2; : ; yn) is speaking to the esteem the class quality Y may hold. Next, the class property estimation of occurrence X could be evaluated by means of computing

$$\arg max\gamma_i \mathrm{P}(\gamma_\mathrm{i}|\mathrm{X}) \tag{1.2}$$

Depending on Bayesian theorem,

$$\mathrm{P}(\gamma_i|\mathrm{X}) = \frac{\mathrm{P}(\mathrm{X}|\gamma_i)}{\mathrm{P(X)}} \tag{1.3}$$

It is important to take under consideration that p(x) is consistent and impartial of yi, thus, we can dismiss the denominator of eq. 1.3 whilst maximizing eq. 1.2. the naive Bayesian classifier presumes conditional independence for making the computations simpler; that is, taking under consideration the class characteristic cost, different characteristic attributes come to be conditionally impartial. this situation, even though unrealistic, performs nicely in exercise and substantially simplifies calculation.

1.5.2 Support Vector Machine

Support vector machines (SVMs) might be considered as a blend of instance-based learning and linear modeling in a high-dimensional space. SVMs could be used for tasks in the case when the data could not be separated via line. SVMs utilize non-linear mapping it transformations instance space to some other space that is of higher dimensionality than original space. In such instance line in new space could be signified as linear boundary in instance space. Originally, the SVMs have been created for the problems related to classification. The concept of Kernel brought upon the SVMs. Kernel might be defined as a function that achieve mapping of a non-linear data to a new space [22].

The Kernel function K is an inner multiplication Φ(x) ● Φ(y) between images of 2 data points x and y:

$$\mathrm{K}(X, \mathcal{Y}) = \Phi(\mathrm{x}) \bullet \Phi(\mathrm{y}) \qquad (1.4)$$

The feature, which kernel function will be expressed as inner multiplication, offers a chance for replacing scalar product with certain kernel choice [23].

The problem of identifying SVM parameters resembles a convex optimizing task, that indicates that local solution is also a global optimum. Separating the data into testing and training sets are involved typically in the classification task. Every one of the instances of the training set comprises single “target value” (class labels) and more than a few “attributes” (features or observed variables). The main aim of SVMs is producing a model (according to the training data) that predict the target values related to testing data having attributes of test data only. SVM for classification are applied for the purpose of finding linear model related to the following form:

$$y(x) = w^T x + b \qquad (1.5)$$

x can be defined as the input vector, while w and b can be defined as the parameters that could be modified for a model and determined in an empirical approach. Concerning the simple linear classification, the task is minimizing a regularized function error which is specified via Eq. 1.6.

$$\mathrm{C} \sum_{n=1}^{N} \xi n + \frac{1}{2} ||w||^2 \qquad (1.6)$$

$$y(w^T\ x + b) \geq 1 - \xi n \tag{1.7}$$

Figure 1.3 shows a sample of linear SVM which was trained on examples from 2 classes. At this point, SVM creates a separating hyperplane, after that it attempts on maximizing the “margin" between the 2 classes. For the aim of calculating the margin, SVM creates 2 parallel hyper-planes, one on every side of the initial one. After that, the hyper-planes will be “pushed” in a perpendicular way away from one another up to the point where they meet the nearest samples from either one of the classes. Such examples are identified as support vectors and are showed in bold in Figure 1.3.

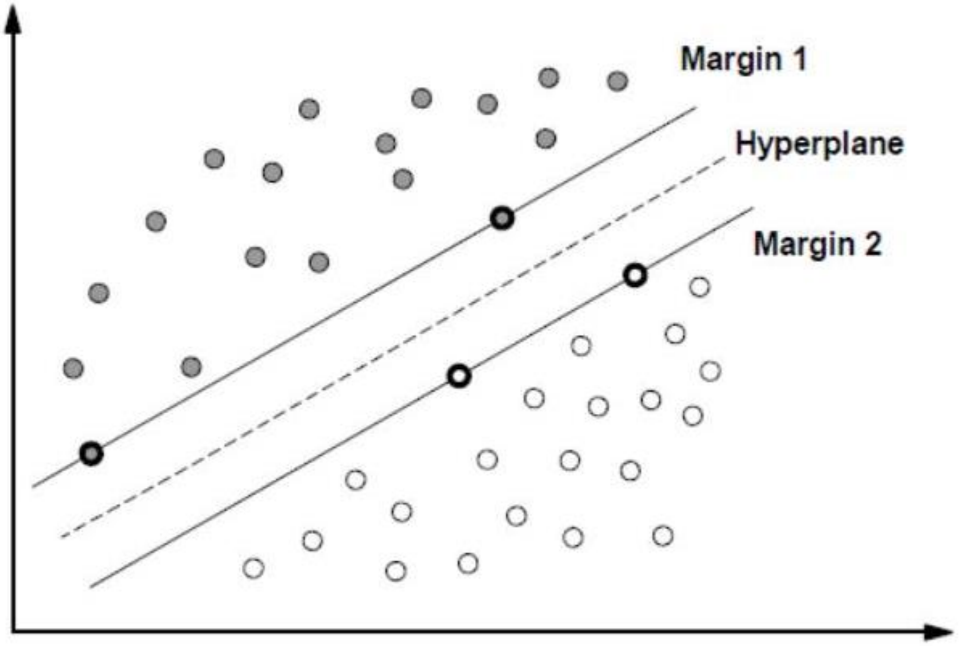


Figure 1.3: SVM Classification

1.5.3 Decision Trees

The major purpose of decision-tree is to be utilized in characterization mechanisms which produce tree structures where each node indicates a test on a trait worth and each branch speak to a test result. The leaves of the tree speak to the classes. The figure shows the choice tree assessed from our Training data set used as part of the project. It demonstrates the connections located in the Training data set. This approach is quick unless the preparation information is substantial. It doesn't give any suspicions about the probability distribution of that data. The process of constructing the tree is called induction [29].

$$E(S) = -\,_{P}(\text{Positive})\ \log 2\ _{P}(\text{Positive}) - \,_{P}(\text{Negative})\ \log 2\ _{P}(\text{Negative}) - \,_{P}(\text{Neutral})\ \log 2\ _{P}(\text{Neutral}) \qquad (1.8)$$

In the expression: p is the probability terms of an event. Entropy is the measure of homogeneity regarding a group of examples. Depending on a group of neutral, negative and positive examples of (the 3-class problem), the entropy regarding set S relation to this binary classification is:

At each tree node, C4.5 choose data attribute that most efficiently divide its sample set into sub-sets that are enriched in one class or other. The criterion of dividing is the normalized information gain (entropy difference). The decision will be made by attribute with maximum normalized information gain. After that, the C4.5 algorithm recurs on smaller subsists. There are some base cases regarding the algorithm:

- All the list samples will be a part of the same class. After that, a leaf node will be created simply for the decision-tree saying for choosing that class.
- Instance of previously-unseen class encountered. Once more, a decision tree will be generated via C4.5 that will be higher up the tree utilizing the probable value.
- Information gain will not be given by any feature. In this instance, a decision node will be generated via C4.5 that will be higher up the tree with the use of the probable class value.

Pseudocode

```
Input: an attribute-valued data-set $D$
    If $D$ is “pure” OR other stopping criterion is satisfied then
        end
    endif
    for each attribute $\alpha \in D$ do
        Calculate information-theoretic criteria if we split on a
    endfor
    $\alpha_{best}$ = Optimal attribute based on criteria that has been calculated above
    $Tree_v$ = Make a decision node which test a best in the root
    $D_v$ = Induced sub-data-sets from $D$ based on an optimal
    for each $D_v$ do
        $Tree_v$ = C4.5 ($D_v$)
```

In data mining step, users’ feelings in a specific tweet is determined with classification tree. Sentiment analysis requires two steps such as learning of a classification model and testing of the classifier. In learning phase, the classification model is trained according to an algorithm. In test phase, unclassified tweets are classified, and predicted classes are compared with real ones. The tweets are evaluated, and class labels are assigned after the text is interpreted as being either positive, negative or neutral.

1.5.4 Random Forests

Ensemble learning concentrates on methods for combining the outputs of various trained models for producing a classifier with extra accuracy. Generally, ensemble models have significantly better-quality performance than the performance of singular model. Random forest algorithm can be considered as an example of an ensemble approach that has been introduced by, it is uncomplicated algorithm, however, in spite of its uncomplicatedness it could offer excellent-quality performance with respect to classification. Figure 1.4 show the main structure of random forest [25].

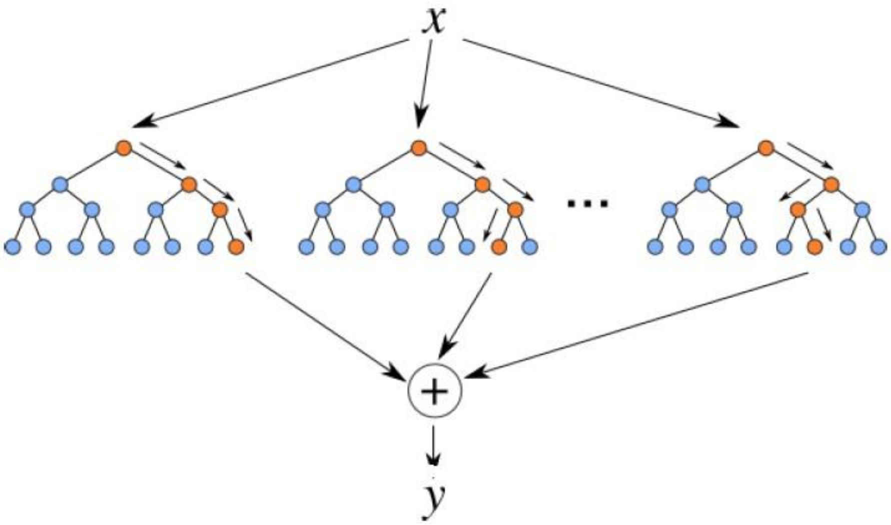


Figure 1.4: The structure of random forest [26]

Combining several decision tree classifiers are utilized for constructing random forests, bootstrapped subset of training data is used to train each tree. Random sub-set of features is selected at every one of the decision nodes, and just splits on these features are considered via the algorithm. The major issue with utilizing a tree lies in the fact that it has high variance that the arrangement of features and training data could have an effect on its performance. All individual trees have high variance, however, in the case when we average over group of trees, we have the ability of reducing the variance of general classification. If every single one of the trees has higher precision then pure chance, and they aren't extremely correlated with each other, the central limit theory indicates the fact that when they're averaged, they are going to create a Gauss distribution. The more decisions which are averaged the lower the variance becomes.

1.6 EVALUATION MEASURES

For evaluating the classification results, some recognized measures from information retrieval will be addressed. All the evaluation measures which are offered in the presented section are based on certain basic counts on testing data collection [27].

The basic measures are counts of false positives (FP), true negatives (TN) true positives (TP), and false negatives (FN) about every one of the classes c of every instance. Those rely on if the class that has been predicted via the classifier match the projected prediction (true class) as displayed in Table 2.1.

Table2.1: Confusion matrix of actual and predicted class.

	C	**¬ C**
C	TP	FP
¬ C	FN	TN

FN, FP, TN and TP will be defined, to signify the number of the respective events took place in group for a class c. The evaluation measures will be defined depending on count statistics. Accuracy (Acc) is considered as the most basic measure. At this point, we measure the ratio related to the accurately classified samples on group D (Eq. 2.8)

.

$$Acc = \frac{\sum_{c \epsilon C} TP}{|D|} \qquad (2.8)$$

In such a case, we have single instance in collection, the accuracy could be estimated via the simplified equation (Equation 2.9).

$$Acc = \frac{TP+TN}{TP+FP+FN+TN} \qquad (2.9)$$

Accuracy is considered as an efficient measure in the case where the classes are scattered in a uniform manner in the set. On the other hand, while class imbalances get more pronounced, a higher level of accuracy can be obtained via a classifier which is biased in the direction of the majority class. Recall and precision are usually utilized as a substitute, thereby giving a more thorough analysis of the behavior of the classifier concerning every class c. Precision P is a measure of the relative frequency of accurately classified samples which have been predicted to be part of c (Eq. 2.10):

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{2.10}$$

Recall (R) is a measure of the relative frequency of accurately classified samples amongst the group of samples whose accurate class is c (Eq. 2.11):

$$R = \frac{TP}{TP+FN} \tag{2.11}$$

The harmonic average of recall and precision is referred to as the F-measure. In the present study, we have utilized the balanced F-measure, or F1 measure, in other words, precision and recall are equally weighted (Equation 2.12):

$$F_1 = \frac{2*P*R}{P+R} \tag{2.12}$$

As opposed to the arithmetic average where Precision = 0 and Recall = 1 (or vice versa), the harmonic average would be 0, and the arithmetic average would be = 0.5. The harmonic average is, in all the cases, less than or identical to the arithmetic average and the geometrical average. In the case where the values of the precision and recall are considerably different, the harmonic average is closer to their minimum rather than their arithmetic average. The measures were suggested, and recommendations have been made by various authors. Based on the findings of some scholars, the harmonic average is of a higher significance as a measure due to the fact that F1 is equal to 1.0 when each of the precision and the recall is perfect and converges to 0 when recall or precision are poor [16], [28], [29], [30],[31].

1.7 THESIS WORK

This research will attention on making use of evaluate the overall performance of diverse classification algorithms regarding this problem. Tweets about computer and models are analyzed in line with the customers' evaluations. Tweets approximately laptop models supply brief feedbacks for companies how to prepare for the unexpected conditions. Pc businesses can start making use of actual-time advertising standards to their campaigns and every day consumer conduct. So, it has big significance several the community of social media users. Many users talk about computer with their emotions. The category gear with (nlp) may be applied to method twitter content. Training information is wanted to be accumulated and pre-processed first, and the class model mastering system is completed next. the schooling dataset is accrued from twitter first by means of the use of MATLAB. Inside the pre-processing step tweets are categorized as tremendous, negative or impartial in line with the records professionals' opinion. Pre-processing step additionally consists of natural language processing strategies such as cleansing, tokenization, and stemming, and so on. The education dataset includes 3000 tweets for iamb.

1.8 REPORT OUTLINE

Introduction in chapter 1, In chapter 2 related work and literature is presented to show the previously adopted approaches to implement data mining and sentiment analysis for twitter contents. Chapter 3 present methodology and implementation of the proposed framework. Chapter 4 present results and performance of the study. After that, conclusion is given, and future studies are detailed at the end of chapter 5.

2. LITERATURE REVIEW AND RELATED WORK

Many researches have been done in text mining area to determine the sentiments in unstructured texts. We will mainly focus on the related works regarding polarity classification with the same objective tasks. Major methods classify given text's polarity at either document, paragraph, sentence, or word levels.

2.1 TEXT MINING

It is a domain used to use statistics mining methods over the textual content; some quantity of early efforts of analyzing exploratory statistics through textual content [32].

Reading sentiments aim for extracting and identifying attitudes, moods, and evaluations of communities and individuals [33].

Presented an early paintings and technical study concerning studying sentiments. at the same time as sentiment analysis and text mining techniques are used collectively within the identical undertaking regarding social media statistics, outcomes are maximum of the times predictive tool or effective descriptive [34].

Text mining turned into had been applied correctly to get posts from fb for the motive of sentiment category via the Arab spring state of affairs.in addition, it may be stated that textual content mining is an understanding coming across approach from natural languages text. The indicated databases have text documents which might be put to reading procedure through utilizing set of algorithms for finding non-trivial and information patterns in the information. As stated earlier than, the maximum commonplace way for storing statistics is storing as textual content. Due to this large portion, it's far assumed that the text mining holds a larger business future than records mining in [35].

Textual content mining is recognized as multidisciplinary approach this is associated with device getting to know, extracting statistics, categorization and text evaluation, and its miles frequently taken into consideration as an extension to statistics mining [35].

No matter the truth that textual content mining and records mining overlap in their fields, text mining is absolutely diverse to information mining whilst considering the character of records its techniques. The structure or form of statistics is very great for expertise the variations among statistics and text mining. common facts mining strategies targeted at the database with based form [35].

Even as text mining is used for processing unstructured facts, called ‘textual statistics’, then extract it’s statistics in automatic manner. The textual facts aren’t dependent in a manner that is just like characteristic cost database. There are most important demanding situations in the natural textual content’s shape. Textual content information is taken into consideration to be fuzzy and unstructured, consequently it's miles complicated to locate beneficial and valid facts styles through an automated approach [35].

Even though the textual content is taken into consideration unstructured, it does have an implied patterns and grammatical shape [36].

This implicit shape is applied and found in numerous methods of specialized textual content mining. natural language processing (nlp) is related to understanding the center of herbal language, additionally it can be applied in preprocessing textual statistics. trustworthy example concerning the rules which could be induced by means of making use of textual content mining-primarily based learning. Textual content mining can be used and fill those gaps which conventional facts mining has made in database information discovery. yet, because of the non-regularly occurring and ambiguous shape of natural language texts it is not viable to generate non-one-of-a-kind and all-embracing algorithms [36].

That the brought about guidelines are primarily based on steady and small corpus, however it'll wrongly expect the meaning of the phrase ‘financial institution’ as a ‘economic institution’ in sentence like “strolling over the financial institution of the river, found a briefcase with cash.” within the previous instance, ‘financial institution’ consult with ground bounding waters. Those regulations might be improved just so extra cases may be expected accurately. Even though, high accuracy class desires extra training value and it could quick cause fuzzy fashions in 2008 [37].

2.2 SENTIMENT ANALYSIS

Some of the state-of-the-art sentiment evaluation applications that are of relevance to our work are crime surveillance [38].

Structures worried with tracking were reviewed and five critical factors which each monitoring machine should have were proposed [39].

A schema become recommended for analyzing the communication of civilians, sentiments and reactions in responding to terrorist attacks [40].

Advised 2 computational techniques for evaluating social media sentiments and the resulted performance had been better while as compared to regular tactics.

Studying sentiment concerning statistics in social media have been applied for detecting sickness outbreaks [41].

The authors described a plan for extracting the tweets for detecting outbreaks and on the spot caution, via a swine flu massive it confirmed an actual contribution to alert involved stakeholders to take a motion. The records in twitter permits for spatial and geographical evaluation [41].

A scheme changed into installed to visualize one-of-a-kind public sentiments taking benefit of the tweets gathered from countries and counties throughout United Kingdom thru the occasion after prince George's start in 2013 this scheme is especially critical in detecting and monitoring downwards and upwards traits of some product allergies unique location at some stage in a time. [42].

The examiner applied textual content mining tactics for investigating patron attitude in the direction of global brands, the studies counseled that twitter could be applied as a valid manner to research attitude towards worldwide brands. Pharmacovigilance is of a significance to our take a look at, it's miles the practice of tracking the outcomes of medical drugs when they were licensed for use [43].

Internet customers should provide on the spot indication approximately harmful drug events from the log date they get even as browsing the net. nearly identical have a look at determined [44].

Where visualization of chfpatients.com discussion board chat sentiments had been applied for measuring the efficiency of the drug by means of quantifying the side consequences of the drug, in particular for the hobby of physicians and forum participants. The goal here is making use of the equal techniques to pills and cosmetics product customers, that is the destined beneficiaries are enforcement and regulatory groups, manufacturers and product users. Analyzing sentiments might be carried out via utilizing two processes. the first method is opinions lexicon-depends on approach [45].

The lexicon consists of a collection of negative and high-quality opinion phrases, the ones phrases are applied to score the opinion sentences with the intention to is both impartial, negative, or fantastic. This approach is very favored and need a scoring function to score each sentence primarily based at the presence negative or positive phrases. the lexicon-based totally sentiment evaluation approach is tested. lexicon-primarily based approach makes use of a lexicon, a group of terrible and fine phrases alongside a scoring feature for the cause of figuring out the sentiment polarity. the second approach is to make use of machine getting to know mechanisms to teach a classifier through utilizing a collection of pre-classifies critiques as a training set. after that the educated classifier is used for classifying sparkling critiques as neutral, terrible, or fine [46].

Utilized a managed mechanism for constructing a classifier making use of n-gram and part of speech tagger approaches and additionally carried out a classifier for evaluations class. the study emphasized that lexicon-based totally strategies surpass machine gaining knowledge of approaches. In 2015 this take a look at makes use of lexicon-based totally processes for analyzing sentiments [47].

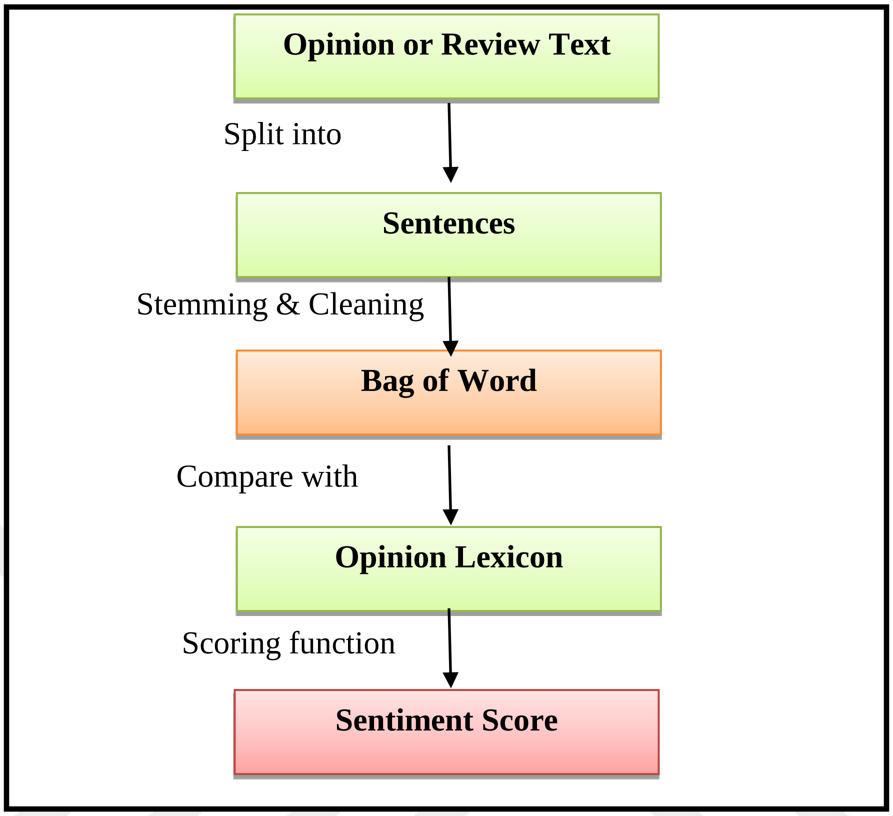


Figure 2.1: Lexicon-Based Sentiment Analysis Approach

The sentiment analysis procedure that is discussed in this have a look at is evaluations-based and its miles often known as opinion mining. whilst most effective sentiment analysis is noted for the rest of our examine, then we suggest evaluations-based totally sentiments analysis. the major aim for reading sentiments is to define the behavior of the opinion holder concerning a count number. Different packages attempt to define the inclusive sentiments in the report. sentiment evaluation might be complex, that is why liu discussed "the trouble of sentiment evaluation". As an example, for that, a textual content may include two or extra critiques concerning numerous or the same items. check the example below in [48].

Example *"I purchased a Nokia Lumia 800 from Amazon. When I unpacked the Nokia phone, I discovered what type of outstanding phone I purchased. The Windows Phone OS is so simple-to-use. Nonetheless as with most phones, the battery life is dramatic. Yesterday, I showed the Nokia phone to my friend, and he immediately became eager. He said that the Nokia Lumia 800 is certainly better than his iPhone 4’’.*

In 2013 the authors utilized textual content mining mechanisms to study client behavior toward worldwide brands, the results confirmed that twitter can be utilized as a dependable method in studying behaviors in the direction of global manufacturers [42].

2.3 PART-OF-SPEECH

Applying grammatical part-of-speech form highlights won't be beneficial for reading sentiments in micro-blogging scope [49].

Analyzing sentiment manner is associated with several observe fields consisting of; analyzing texts, computational linguistics, and nlp. It refers to extracting subjective information from uncooked information, most of the instances in textual content form. yet, other forms of media can have subjective records, inclusive of films, sounds, and snap shots, but some of these sorts studied in smaller extent. In accordance, numerous kinds of sentiments exist in in all forms of media. Sentiments could discuss with emotions or opinions. even though critiques and feelings are related to any other, however, there's an apparent distinction among them [50].

In opinion-primarily based sentiment evaluation, a distinction is present among poor and superb reviews. even as in feelings-primarily based sentiment analysis, there may be difference between various emotion kinds which includes being unhappy, happy or angry, happy, unhappy etc. Due to the dynamic, heterogenous, big and multisource capabilities of the utility information utilized in a allotted environment, one good sized function regarding massive records is wearing out calculate in petabyte and exabyte stages facts with a ganglion calculation manner. consequently, applying parallel computing framework, it is corresponding programming language assist, additionally software program models that successfully mine and examine allotted information are widespread goals in large information processing for the reason of converting from amount to quality.

2.4 MAPREDUCE

With reading these conventional algorithms concerned with device learning, we kingdom that the computational strategies in set of rules' gaining knowledge of method can be transformation to a summation operation on some of training statistics units. The summation system can be implemented on diverse sub-units one at a time and acquire penalization achieved in clean way on MapReduce programming scheme. Hence, an enormous-scale fact set may be split to some of sub-units and allocated to more than one mapper nodes. After that, extraordinary summation techniques may be applied on mapper nodes for the motive of amassing common consequences. In the long run, gaining knowledge of algorithms are accomplished in parallel manner by means of blending summation on diminish hubs [51].

Accomplished cautioned MapReduce based software programming interface phoenix, that help parallel programming in multiprocessor and multicore gadget environments and identified 3 algorithms concerning records mining such as linear regression, foremost component analysis and okay-manner [52].

Upgraded the implementation mechanism of MapReduce's in Hadoop, additionally they dispensed data garage, tested data sharing between the computing nodes used in parallel getting to know algorithms, expected the overall performance of algorithms concerning query primarily based studying, iterative mastering and single-pass getting to know in MapReduce scheme, then they confirmed that MapReduce strategies are applicable for huge-scale information mining through testing series of standard information mining duties on medium size clusters [53].

Papadimitriou and sun cautioned an allotted collaborative aggregation (disco) scheme by way of utilizing collaborative aggregation and efficient disbursed records pre-processing mechanisms. The software on Hadoop in an open-source MapReduce assignment validated that disco has terrific scalability and can examine and manner big data units (with loads of gb). For the purpose of improving inadequate scalability of not unusual analysis software and weak reading competencies of Hadoop structures [54].

Applied a study regarding the mixing of (open supply statistical evaluation software) and Hadoop. The specified integration push records computation to parallel processing, that permit sturdy precise evaluation skills for Hadoop [55].

Succeeded in attaining integration of weak (open supply records mining and system gaining knowledge of software program device) and MapReduce. Not unusual weak equipment ought to handiest operate on one gadget, with simplest 1-gb memory. Following to algorithm parallelization, weak conquer these obstacles and improve the performance through taking gain of the use of parallel computing to hold more than a hundred gb statistics on MapReduce clusters. Suggested Hadoop-ml, wherein that developers may want to virtually construct facts-parallel or challenge-parallel information mining and device getting to know algorithms on software blocks beneath the language runtime surroundings [56].

2.4 TWITTER SENTIMENT ANALYSIS

Twitter, a microblogging internet site which is these days acquainted to the majority, has made a amazing development in reputation and usage inside the beyond two years. The twitter and its workload exam are isolated into two set: collecting and ad hoc question. in 2008, twitter acquired a full-size morale analysis, which enabled twitter to come across unsolicited mail messages at the platform in [57].

Idea for information mining from twitter must consist of a scalable structure for huge datasets collections. Additionally, to make certain the transfer and don't forget the processing abilities that are available. On the alternative side, ad hoc queries, in keeping with are submitted explicitly via users, however those queries have not any computational pattern. As task scheduler is carried out workflow supervisor, oink that ambitions to take care of dependencies among submitted software jobs [57].

In step they have got used apache Hadoop with unmarried node cluster, apache pig, hdfs and selection tree. One a part of device studying is sentiment evaluation and can be applied with twitter to get an analysis of the evaluations of twitter customers about products, vital political event, election consequences, brands and their function in the marketplace. after reading the evaluations of clients to achieve the results of advantageous or bad sentiment [57].

To twitter sentiment analysis there are ultra-modern approach is based on the researched. The researchers cope with challenges to powerful choice of twitter records based totally on hashtag (#). Information training also used a set of expression records in addition to records acquisition with (high-quality, bad and impartial) emotions for assessment functions [58].

In which the pre-processing phase changed into divided into tokenization, normalization and part-of speech (pos). It’s far a critical part of the linguistic pipeline and the principle form of grammatical evaluation. twitter sentences can be constructed from a noun, verb, adjective or different components of speech [50].

It's miles the manner of identifying symbols and abbreviations, which include brb, forestall phrases or other running a blog capability. more precisely, normalization is the system of changing abbreviations into their proper which means, an instance of that's "be right returned" (brb). Furthermore, the uppercase characters in the twitter content material had been changed to decrease case. the hunt results [58].

Confirmed that part of the speech indicates that the wide variety of appearance nouns, verbs etc. and therefore, haven't any potential for twitter sentiment analysis. but, has verified the significance of the hashtag and emotions method of the sentiment evaluation. Some other method is to twitter sentiment analysis examine. Classification of tweets had been constructed on a version with primary capabilities. binary tweets are classified according to sentiment in high quality and bad units handiest, and there's additionally a third way of classifying impartial emotions. the pre-processing phase was [59].

Which became proposed with a dictionary of expressions used to label all expressions with a statement of feelings and a quick dictionary, which have been translated into acronyms in English, consisting of “br8k” into “wreck”. Also, the URLs have been protected in twitter and replaced together with (@ john) with the tag in which all the negations (as an instance: not, not, can't) were changed with “now not” and all the repeated word characters (for example: cooool) also are shortened to (cool). Research has shown that the evaluation function offers the exceptional effects accurately by way of combining polarity before the words and tags of the pos. The research discusses the implicit conclusions of a style of linguistic-like phrases in twitter [60].

The semantic styles are defined as the group of phrases in corporations with consultant sentiment semicircle model [60].

Represents the process of extracting sentiment styles from tweets, parent 3. The tweeters are accumulated geologically prior to processing for the purpose of noise filtering. after that, the lexicon of sentiment is applied to gather contextual connotations between twitter words. Then the same words are grouped inside the tweets the cause of that is to form the semantic patterns.

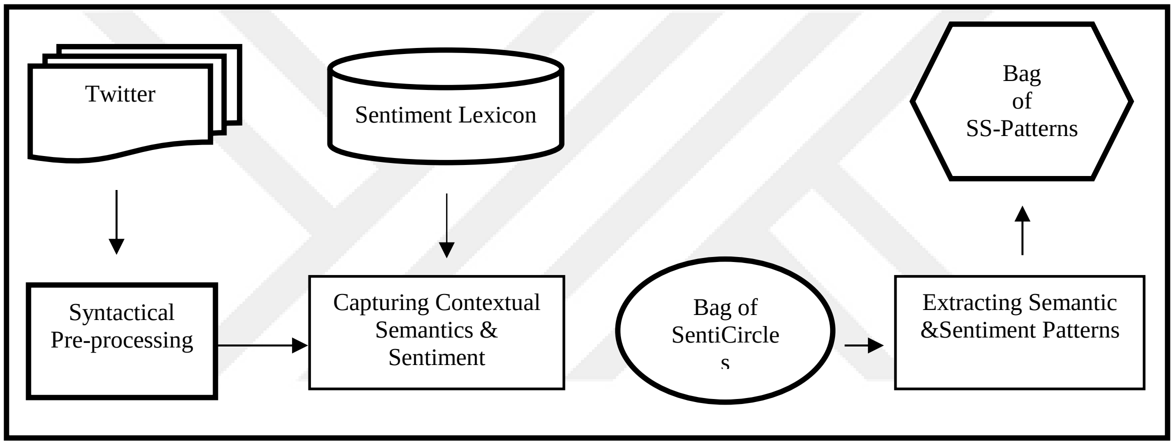


Figure 2.2 The process of extracting patterns from Twitter data [60].

The effects of researcher led to the class of sentiments of the entity that effective sentiments were simpler to search for tweets, in place of bad or impartial. the purpose for that is that the selected classifier or range of tweets may be affected with positive sentiment in the statistics set. In which social media and social networks entered into the examiner that offers with imaginary studies [60].

The feasibility observes on whether tweets generated by way of twitter users talk over with unique and specific subjects that may exceed marketplace-based totally revenue forecasts approximately the contemporary films. To discover people's evaluations and behavior about extraordinary films, sentiment analysis was used. The not unusual sentiment between "fans" screen a first-rate deal of sales, and the consumer could have an influential opinion approximately new and unique film commercials. films are normally shown on Fridays, aside from some that are shown on Wednesday. to achieve consistency wherein the films that have been launched on Friday were only considered too were launched on a huge scale. For those films that have been confined edition, we started gathering statistics from the time it was searching wide. the to be had dataset incorporates the tweets that had been requested via the twitter seek abe, where the search is done by using the identity of the film. the primary attributes of each tweet had been the writer, textual content and timestamp. furthermore, the researchers carried out a one-week evaluation of the consumer's feelings earlier than the release of the brand-new film, wherein all promotions together with pics and trailers are set [61].

The solution has been evolved with a linear regression version that ends in predicting or predicting film sales before the dates in their statement. The famous dating between users and twitter related subjects in a specific film was announced and its success in the future [61].

Stock marketplace prediction evaluation based on twitter is researched in j. aim is to predict assertion, if collective mood extracted from tweets is correlated with down jones commercial common index over time frame. behavioral economics define that human choice-making can be influenced by our emotions or mood settings [62].

Opinion finder equipment changed into deployed in technique to analyses tweets by means of poor or effective values and google profile of mood states (gloms) to research person temper from tweets that degree six awesome mood dimensions [62].

Moreover, self-organizing fuzzy neural network turned into applied to make testing whether accuracy of prediction can be more desirable via temper measured of twitter customers. consequences confirmed that not all mood dimensions ought to correlate to, which means adjustments of public mood within dimensions can match adjustments in figures that occur later up to four days. twitter as social network offers skills for predicting numerous situations, along with mentioned earlier than: correlations between twitter and movie sales or inventory [62].

2.5 ANALYTIC OF TEXT MINING METHODS

With the developing significance of social networks and of blogs, sentiment analysis and opinion mining have become a great area of examine for many researches [33].

An extensive evaluate of the modern observe become cautioned. of their observe, they described the cutting-edge strategies and strategies for retrieving opinion-orientated data. yet, now not so many researches within the opinion mining discipline positioned blogs into attention or even an awful lot less researchers taken into consideration microblogging. the authors in c. yang et al. in 2007 [63].

Applied internet blogs to build a corpus for analyzing sentiments and utilize emotion icons with blog posts as a way for representing the mood of the user. authors used crf and svm newcomers for classifying sentiments at sentence degree, then they examined some of mechanisms for the reason of determining the general sentiments in the record. therefore, the great mechanism is recognized by thinking about the sentiment of the last sentence of the report as the sentiment on the file stage [64].

Utilized emoticons like “:-)” and “: - (” to construct a training set for classifying sentiments. because of this, the authors accumulated a few texts inclusive of emoticons from UseNet newsgroups. the accumulated information-set have been divided into classes “poor” (textual content with angry or unhappy emoticons) and “high-quality” (text with happy emoticons) samples. emoticons skilled classifiers: naïve Bayes and svm, were successful to stand up to 70% of accuracy percent concerning the trying out set [65].

Utilized twitter for accumulating schooling statistics, then they to implement a sentiment seek the authors went for a similar approach to construct corpora through utilizing emoticons to get “terrible” or “advantageous” samples, after that applying exceptional classifiers. the best result changed into acquired by naive Bayes classifier with a mutual statistics degree for function choice. The researchers had been successful of getting as much as eighty-one % accuracy percentage on their trying out set. yet, the technique presented a weak overall performance with three instructions (“impartial”, “fantastic” and “poor”) [64].

3. MATERIAL AND METHODS

Data collection is a difficult task and it is not easy. We must determine the specific path to obtaining the data for the thesis, and for my thesis I will choose the training data and emotion analysis from the Blog. In this chapter we will talk about collecting, processing, storing and analyzing tweets.

3.1 METHODOLOGY

To acquire this objective mentioned on top the subsequent purposed is used:

∑ An extreme investigation of existing systems and techniques in order of assumption examination.

∑ Arrangement of associated measurements from twitter with the need of tpi.

∑ Handling of records takes from twitter all together that it ought to be fit mining.

∑ To make a class cation upheld unmistakable managed gadget discovering ways.

∑ Instructing and endeavoring out of assemble classifier the use of immense datasets

∑ Processing the consequences of very surprising classifier, the utilization of dataset expanded from twitter.

3.2 PROPOSED OF ARCHITECTURE

As we will probably acquire conclusion investigation for data provided from twitter. We are going to develop a classifier which is made out of different AI to know classifiers.

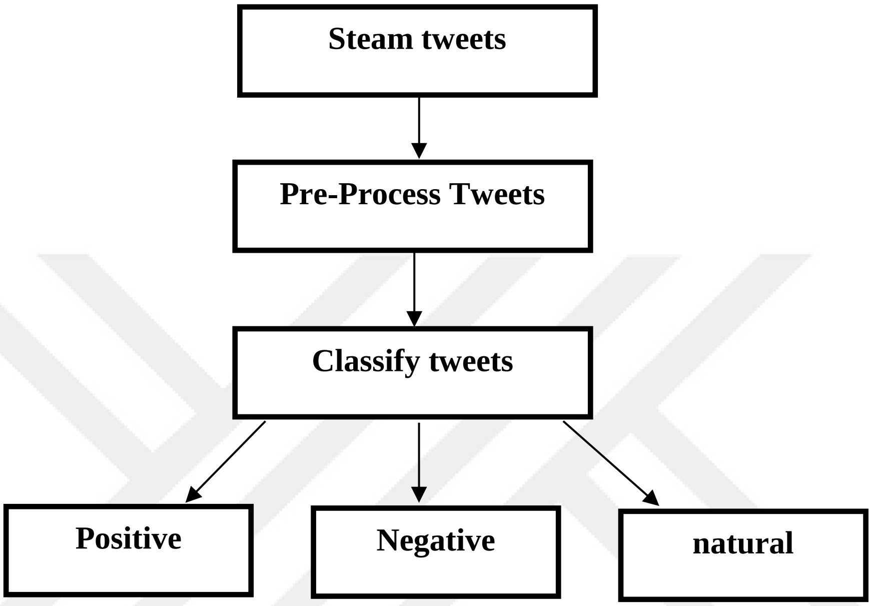


Figure 3.1: classify tweets of twitter using classifier technique

1- First, we're going to stream tweets in us develop classifier with the help of tweepy library in MATLAB.

2. Then we are preparing those tweets, all together that they might be suit for mining and trademark extraction.

3. After preparing, the information goes through the preparation classifier of the work that partitions it into positive, negative and nonpartisan.

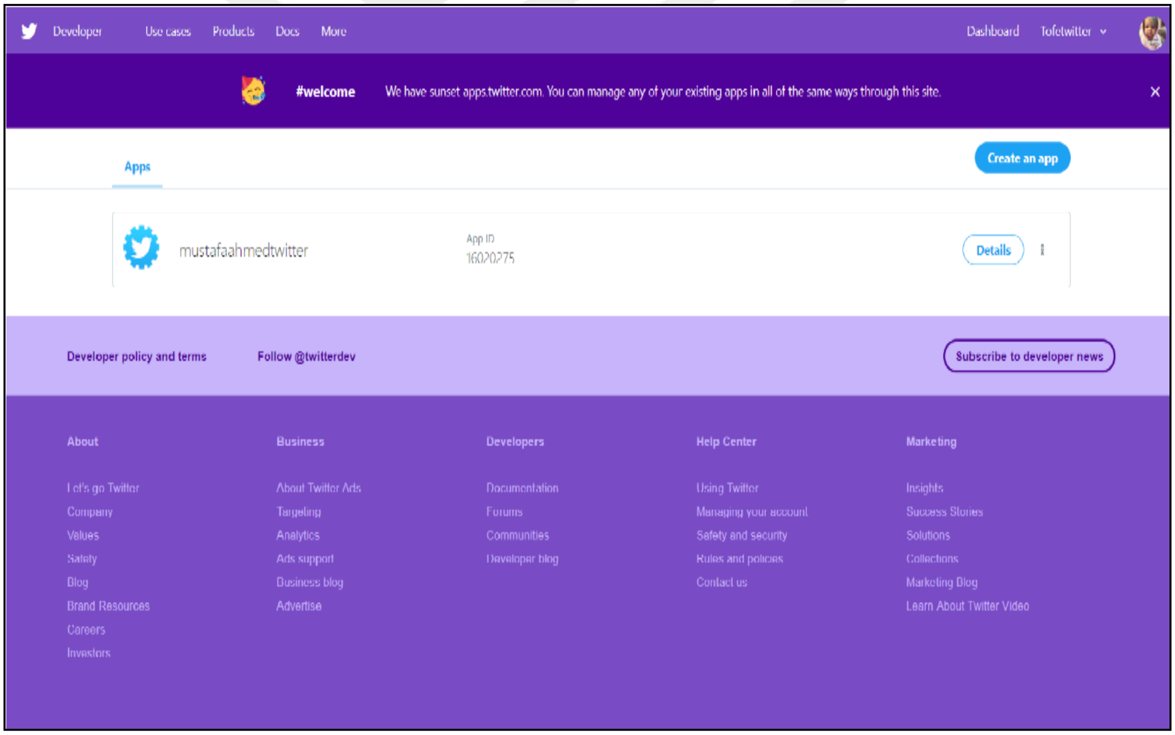
Given that, twitter is our wellspring of data for assessment. We are going to move the tweets from twitter in our database. For this we're going to utilize twitter blog.

3.3 TWITTER API (APPLICATION PROGRAMMING INTERFACE)

Application Programming Interface (API) is an interface that enables interaction with web services. It gives developers and public to develop service products on top of API and thus implement it within own service solutions. Access to Twitter service is possible by two types of different methods: Streaming API and REST API.

3.3.1 Obtaining of Twitter Application

Firstly, the manner of having access token starts with acquiring new application from twitter utility control panel by using filling up wanted attributes for application, such as: name, description and website.



#welcome We have sunset apps.twitter.com. You can manage any of your existing apps in all of the same ways through this site.

mustafaahmedtwitter

App ID
16020275

Figure 3.2: Twitter Application Setup

3.3.2 Obtaining Twitter Credentials

For the development and streaming api access are essential credentials placed below keys and get right of entry to tokens tab in newly generated twitter mining11 utility settings. there are 4 credentials to observe client secret api secret and patron key api key get admission to token mystery and get right of entry to token. those credentials provide the entirety for twitter mining11 utility to authorize itself and make api requests on its proprietor mustafaahmedtwitter behalf. Owner username mustafaahmedtwitter represents researchers own twitter account. Configuration control panel with twitter utility settings. Credentials are blacked as they present touchy fact.

4. EXPERIMENTS AND DISSCUSION

In this section, the experiment will be offered beside with the outcome and the calculation, separated into chapter as the follows:

4.1 MATLAB TOOL

MATLAB is a user interface enterprise instrument projected exactly for engineers and academics? The sensitivity of MATLAB is the MATLAB language, a matrix-based programming language authorizing the uppermost normal onset of computational mathematics.

- Data study
- Algorithms grow
- Generate requests and replicas

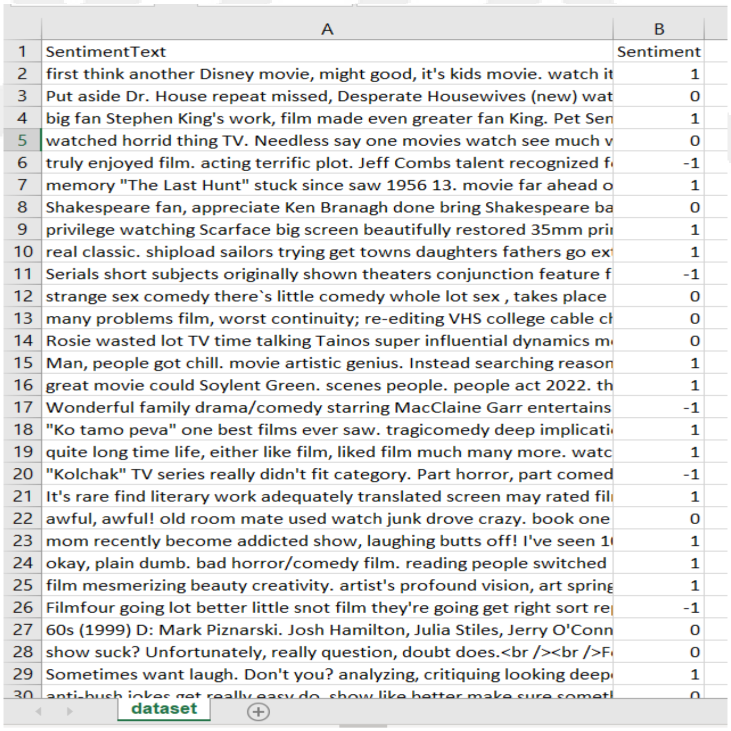
The language, apps, and built-in math purposes permit the scientist to quickly learn various approaches to spread at an answer. MATLAB leases the specialists produce your opinions from study to production by putting to inventiveness applications and rooted plans, as well as fraternization with Stimulant and Model Founded Design.

4.2 DATA COLLECTION

First, you must create a Twitter account to be able to use the Twitter API. It is not only difficult to fill out information on the Twitter site and the company will provide you with a password and a username by logging on to Twitter. You can read the tweets and send and receive comments on any topic you want. As for how to get a developer account, through your own account you can log on to Twitter developers by creating an account to write tweets through some important details g After the creation of a developer account can be used client key, the client secret key, access to the code key and access to the secret key in Access to Tweets.

4.2.1 Tweets Collected

The tweets were obtained by the Twitter API being stored in a csv file shown in Figure 4.1



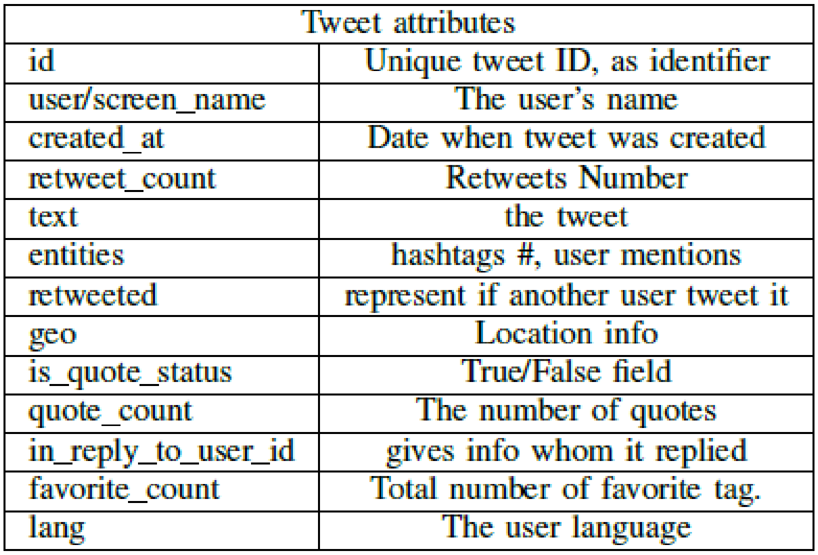
	A	B
1	SentimentText	Sentiment
2	first think another Disney movie, might good, it's kids movie. watch it	1
3	Put aside Dr. House repeat missed, Desperate Housewives (new) wat	0
4	big fan Stephen King's work, film made even greater fan King. Pet Sen	1
5	watched horrid thing TV. Needless say one movies watch see much v	0
6	truly enjoyed film. acting terrific plot. Jeff Combs talent recognized f	-1
7	memory "The Last Hunt" stuck since saw 1956 13. movie far ahead o	1
8	Shakespeare fan, appreciate Ken Branagh done bring Shakespeare ba	0
9	privilege watching Scarface big screen beautifully restored 35mm pri	1
10	real classic. shipload sailors trying get towns daughters fathers go ex	1
11	Serials short subjects originally shown theaters conjunction feature f	-1
12	strange sex comedy there`s little comedy whole lot sex , takes place	0
13	many problems film, worst continuity; re-editing VHS college cable ch	0
14	Rosie wasted lot TV time talking Tainos super influential dynamics m	0
15	Man, people got chill. movie artistic genius. Instead searching reason	1
16	great movie could Soylent Green. scenes people. people act 2022. th	1
17	Wonderful family drama/comedy starring MacClaine Garr entertains	-1
18	"Ko tamo peva" one best films ever saw. tragicomedy deep implicati	1
19	quite long time life, either like film, liked film much many more. watc	1
20	"Kolchak" TV series really didn't fit category. Part horror, part comed	-1
21	It's rare find literary work adequately translated screen may rated fil	1
22	awful, awful! old room mate used watch junk drove crazy. book one	0
23	mom recently become addicted show, laughing butts off! I've seen 1	1
24	okay, plain dumb. bad horror/comedy film. reading people switched	1
25	film mesmerizing beauty creativity. artist's profound vision, art spring	1
26	Filmfour going lot better little snot film they're going get right sort re	-1
27	60s (1999) D: Mark Piznarski. Josh Hamilton, Julia Stiles, Jerry O'Conn	0
28	show suck? Unfortunately, really question, doubt does.<br /><br />F	0
29	Sometimes want laugh. Don't you? analyzing, critiquing looking deep	1
30	anti-bush jokes get really easy do. show like better make sure someth	0

Figure 4.1: Sample Tweets Collected

4.2.2 Tweets Attributes

The data we obtained is relevant to computer hashtags from twitter, 3000 tweets are gathered completely. the dataset acquired, from twitter has 1828 distinctive. every field explain distinct attributes of that tweet inclusive of retweet reputation, reply to, language, vicinity, consumer information, hashtag details etc. some of them are defined in desk four.

Table 4.1: Different Fields of Twitter Dataset



Tweet attributes	
id	Unique tweet ID, as identifier
user/screen_name	The user's name
created_at	Date when tweet was created
retweet_count	Retweets Number
text	the tweet
entities	hashtags #, user mentions
retweeted	represent if another user tweet it
geo	Location info
is_quote_status	True/False field
quote_count	The number of quotes
in_reply_to_user_id	gives info whom it replied
favorite_count	Total number of favorite tag.
lang	The user language

From those diverse attributes, the textual content subject is the most vital for sentiment evaluation and we pick out the textual content field and erase different needless fields in our dataset as proven in parent 15. Amassing tweets from streaming api are saved in csv file, but there's capability to view near actual-time streaming, that is used for cause in pre-processing technique.

4.3 PRE-PROCESSING OF DATA

4.3.1 Tokenization

Tokenization is a procedure of part message strings into tokens, which are represented by words in sentences. As part of Twitter data analyses, this method will help with Named Entity Recognition. From the tweet dataset we can request sample data and tokenize them to see the results. Purpose of tokenization is to analyses ways of efficient tweet splitting into tokens and comma-separated mentions [75].

Table 4.2: Tokenized Dataset

Text	After process applying
I dont see that Loving for most #Imdb	"I" "dont" "see" "that" " Loving " "for" "most" #Imdb
True power shines through from any perspective the Imdb Concept #5500 Performance	"True" "power" "shines" "through" "from" "any" "perspective" "The" " Imdb " "Concept" “#” "5500" "Performance"

4.3.2 Removing numbers and punctuation

It is recognized that the numbers are utilized for representing measures, year, and financial representation etc. As numbers are not necessary for work analyzation in this step, the numbers must be detracted from the information. Accordingly, numbers should detract from an information.

Table 4.3: Removing Numbers and Punctuation

Text	After process applying
"I" "dont" "see" "that" "Loving" "for""most” #Imdb	"I" "dont" "see" "that" "Loving" "for" "most” #Imdb
"True" "power" "shines" "through" "from" "any" "perspective" "The" " Imdb " "Concept" “**#**” "5500" "Performance"	"True" "power" "shines" "through" "from" "any" "perspective" "The" " Imdb " "Concept" "5500" "Performance"

4.3.3 Changing Uppercase Letters to Lowercase

Words with upper-case or lower-case are counted in a different way in a case sensitive situation. Unfortunately, they share the same meaning in a sentence. Therefore, capital letters are converted to their lower case.

Table 4.4: Changing Uppercase Letters to Lowercase

Text	After process applying
"**I**" "dont" "see" "that" "**Loving**" "for" "most” “Imdb "	"**i**" "dont" "see" "that" "loving" "for" "most” “imdb "
"**True**" "power" "shines" "through" "from" "any" "perspective" "**The**" " Imdb " "**Concept**" "**5500**" "**Performance**"	"**true**" "power" "shines" "through" "from" "any" "perspective" "**the**" " imdb " "**concept**" "**5500**" "**performance**"

4.3.4 Removing the stop-words

Stop words are not valuable words such as pronouns, adverbs, auxiliary words and some conjunction words. Stop words are general words that does not have any meanings. They should be also removed from text in the preprocessing.

Table 4.5: Removing the Stop-Words

Text	After process applying
"**i**" "dont" "see" "**that**" "loving" "**for**" "**most**” “imdb "	"dont" "see" " loving " " imdb "
"true" "power" "shines" "**through**" "**from**" "**any**" "perspective" "**the**" " imdb " "concept" "5500" "performance"	"true" "power" "shines" "perspective" " imdb " "concept" "5500" "performance"

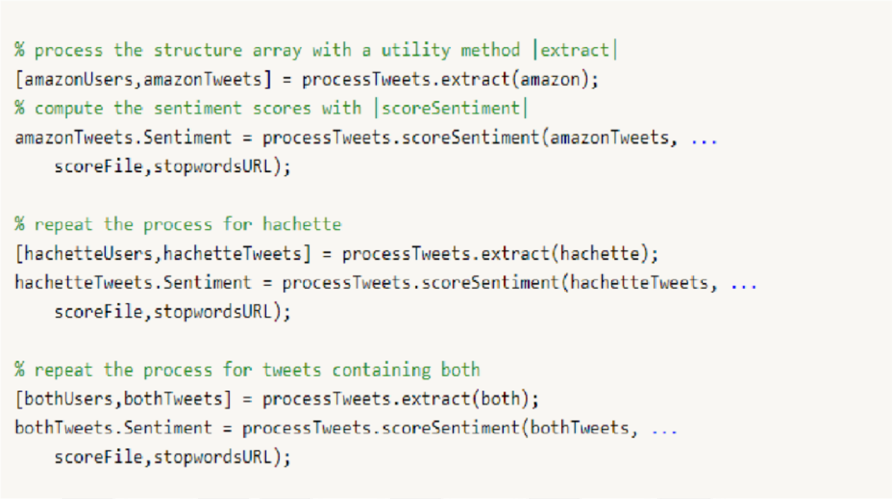
4.3.5 Stemming:

This preprocess step is performed on word roots. If base form of words has the same meaning, then they can be analyzed as one term. Without stemming, they could be assumed as different and unique words. Stemming process is needed to reduce word variants and increase the accuracy.

Table 4.6: Stemming

Text	After process applying
"dont" "see" "**loving**" " imdb " "	"dont" "see" "**love**" " imdb "
"true" "power" "**shines**" "**perspective**" " imdb " "concept" "5500" "**performance**"	"true" "power" "**shine**" "**perspect**" " imdb " "concept" "5500" "**perform**"

Following code shown all the steps in Figure 4.2



```
% process the structure array with a utility method |extract|
[amazonUsers,amazonTweets] = processTweets.extract(amazon);
% compute the sentiment scores with |scoreSentiment|
amazonTweets.Sentiment = processTweets.scoreSentiment(amazonTweets, ...
    scoreFile,stopwordsURL);

% repeat the process for hachette
[hachetteUsers,hachetteTweets] = processTweets.extract(hachette);
hachetteTweets.Sentiment = processTweets.scoreSentiment(hachetteTweets, ...
    scoreFile,stopwordsURL);

% repeat the process for tweets containing both
[bothUsers,bothTweets] = processTweets.extract(both);
bothTweets.Sentiment = processTweets.scoreSentiment(bothTweets, ...
    scoreFile,stopwordsURL);
```

Figure 4.2: Text Preprocessing

4.4 CLASSIFICATION

Deep Sparse Autoencoder from the structural connect of recognize, the autoencoder is an axisymmetric base hit hidden-layer neural join [76].

The autoencoder encodes the input sensor broadcast by per the disoriented layer, approximates the minimum lapse, and obtains the best-feature hidden-layer conceit [77].

The work of genius of the autoencoder comes from the unsupervised computational pose of cro magnon man perceptual training [78].

Which itself has some down-to-earth flaws. For concrete illustration, the autoencoder does not revoke any practical achievement through copying and inputting hallucination into suggested layers, during it bounce reconstruct input message with fancy precision. The rare autoencoder inherits the sense of the autoencoder and introduces the sparse comeuppance term, adding constraints to feat learning for a concise conceit of the input statement [77, 78].

$$p_j = \frac{1}{n} \sum_{i=1}^{n}[a_j\ (x(i))] \tag{4.1}$$

The hidden layer is unbroken at a lower worth to confirm that the typical activation worth of the distributed parameter is outlined as, and therefore the penalty term is employed to stop from deviating from parameter. The Kullback–Leibler (KL) divergenceis employed during this study because the basis of social control. The mathematical expression of KL divergence is as follows:

$$kL(p||p_j) = p\ln\frac{p}{pj} + (1-p)\ln\frac{1-p}{1-p_j} \tag{4.2}$$

When doesn't deviate from parameter, the KL divergence worth is 0 otherwise, the KL divergence worth can bit by bit increase with the deviation? the price operate of the neural network is about as C (W, b).

4.5 MEAN SQUARE ERROR

In measurements, the mean that square blunder (MSE) or mean that square deviation (MSD) of an estimator (of a procedure for approximating an unnoticed sum) occasions the normal of the squares of the mistakes that is, the regular squared change between the assessed standards and what is predictable. The MSE can calculated in equation (4.3).

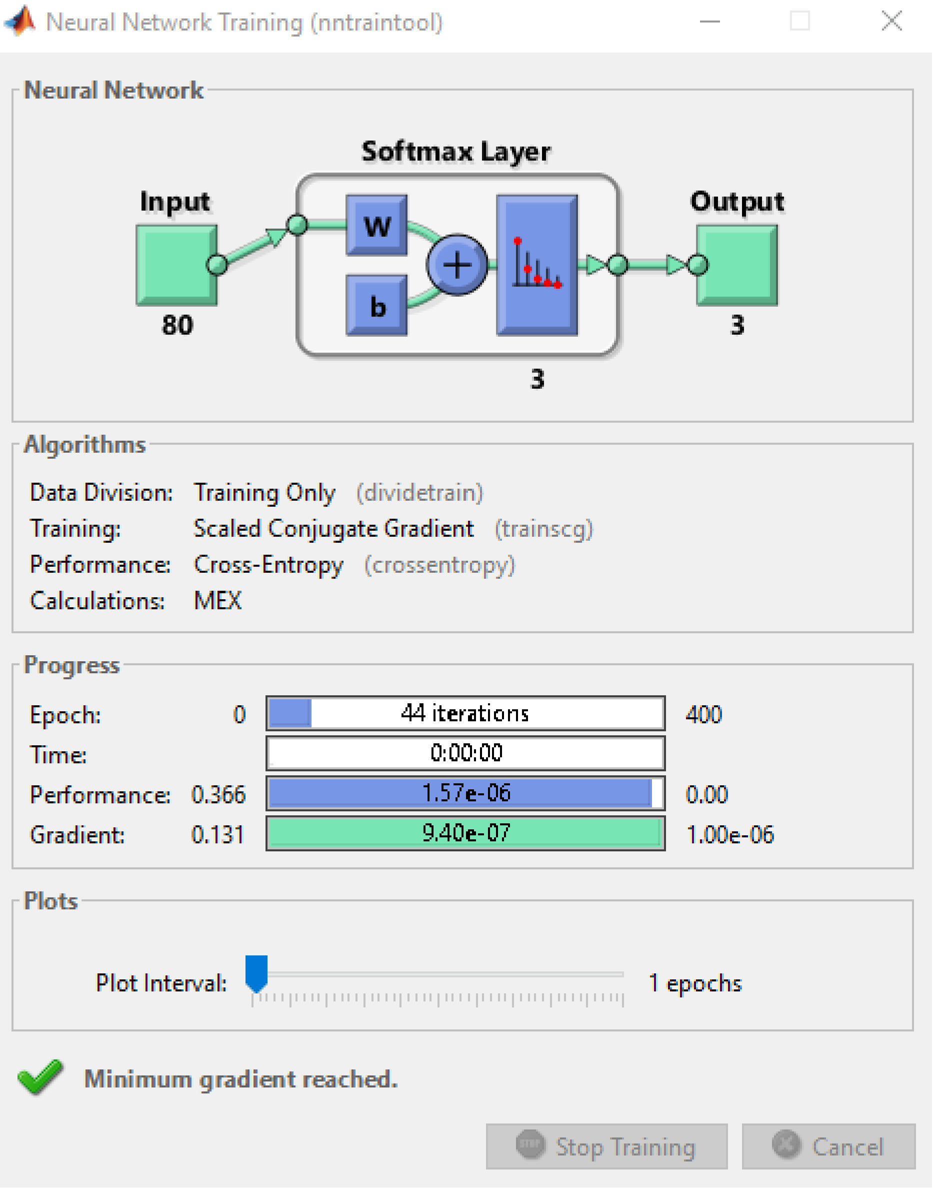
$$\text{MSE} = \frac{1}{m}\sum_{k=1}^{n}(Y_k - F_k)^2 \tag{4.3}$$

Where m is amount of instance, k is the numeral of features F characterized the input features. Y characterized output data.

4.6 IMPLEMENTATION

Numerous experiments are performed in several settings, the test and the outcome were measured applying several measurements, the performance of several experiments were related, and the outcomes were highlighted.

The experimental applied by using MATLAB2018 as tool. The Figure 4.4 display numeral of material like total of epoch, time and performance. Moreover, the construction of the network also offered in this Figure which there is 3 input and 1 output the best essential features.



Algorithms

Data Division: Training Only (dividetrain)
Training: Scaled Conjugate Gradient (trainscg)
Performance: Cross-Entropy (crossentropy)
Calculations: MEX

Progress

Epoch:	0	44 iterations	400
Time:		0:00:00	
Performance:	0.366	1.57e-06	0.00
Gradient:	0.131	9.40e-07	1.00e-06

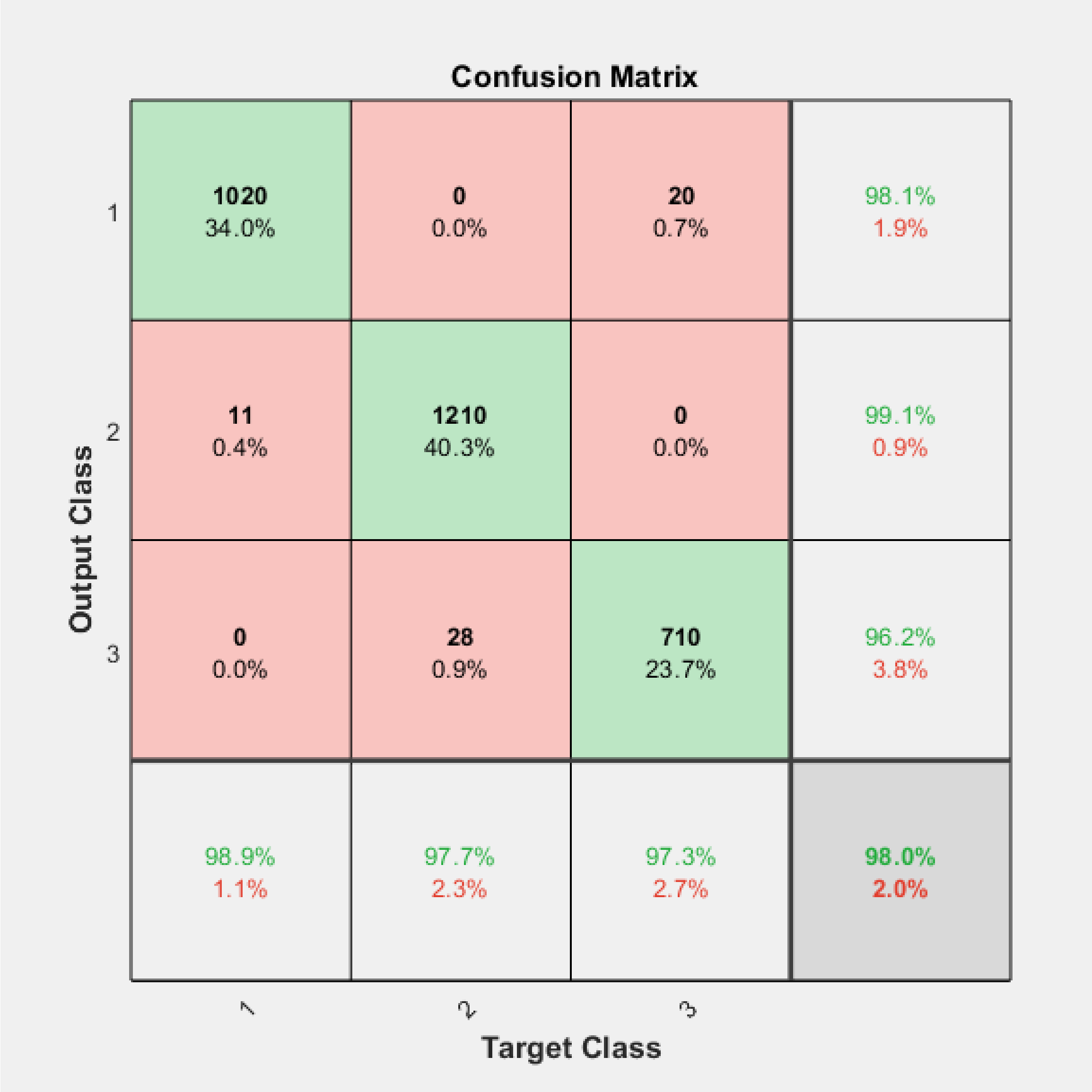
Plots

Plot Interval: 1 epochs

Minimum gradient reached.

Figure 4.3: Deep Sparse Autoencoder Training processes**.**

The confusion matrix is calculated to evaluate the performance of the proposed method see Figure 4.4 10 parameters are calculated to evaluate the performance of the proposed method and presented.



Confusion Matrix

Output Class \ Target Class	1	2	3	
1	1020 34.0%	0 0.0%	20 0.7%	98.1% 1.9%
2	11 0.4%	1210 40.3%	0 0.0%	99.1% 0.9%
3	0 0.0%	28 0.9%	710 23.7%	96.2% 3.8%
	98.9% 1.1%	97.7% 2.3%	97.3% 2.7%	98.0% 2.0%

Figure 4.4: Confusion Matrix.

Table 4.7: Proposed Method Results.

Parameters	**Results**
Sensitivity	0.9600
Specificity	1.0000
Precision	1.0000
Negative Predictive Value	0.9600
False Positive Rate	0.0000
False Discovery Rate	0.0000
False Negative Rate	0.0400
Accuracy	**0.9800**
F1 Score	0.9760
Matthews Correlation Coefficient	0.9680

Maximum Accuracy after applying the classification technique is 0.98. As shown in Table 4.7 the proposed method presented high results when 10 statistical parameters are calculated. Then, the obtained results compared with well-known studies presented in this field. According to the comparison in the Table 4.8 our method presented remarkable results compared to previous studies.

Table 4.8: Results Comparison.

Methods	**Results**
Deep Learning with Bi-LSTM [79]	94
SVM [80]	91
Maximum Entropy algorithm [81]	70.04
Proposed Method	**98.00**

As shown in the Table 4.8 the proposed method presented best results than methods proposed in which these studies represented the commonly known researches in this field [79] [80] [81].

5. CONCLUSION

Opinion mining is a field where a large data volume is being generated via person-to-person communication. Analyzing sentiments is a field used in various applications such as advertising, social media and forensic. By the assistance provided by opinion mining, the corporations could estimate their market and learn which changes considered to be necessarily needed to make for next product up gradation. Furthermore, providing mechanisms to construct a plan on their item. Customers can also use opinion mining for buying a product which they never used before, as the customer always prefer sentiments and reviews associated to the product before they purchase it. First dataset is collected from twitter by using Twitter Api. Twitter streaming API is applied on MATLAB for obtaining 3000 tweets about imdb. These messages are prepressed and cleaned.

Since, Twitter text classification has many challenges to optimize the streaming, different supervised and unsupervised machine learning algorithms can be implemented in order to recognize patterns between tweets and make data driven predictions.

Opinion mining is complex to examine subjective and objective data, a large opinionated word likewise happens in target sentences, thus it is very hard to control those challenges. Usually the users post their surveys in the forums or blogs with a lot of spelling mistakes which our word reference cannot detect them and bringing about less exactness of sought yield. Along these lines, part of work must be done in this field for recognizing spam online, word sense disambiguation.

Moreover, other machine learning algorithms for text classification can be implemented in future work. Comparing their accuracy can lead to enhanced results. Pre-processing approach with data normalization can be extended with natural language processing on the higher level, which could optimize the sentiment classification.

REFERENCES

[1] A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimera, (2004) “Community analysis in social networks,” Eur. Phys. J. B, 38(2), pp. 373–380.

[2] G. Simos, (2015),“How Much Data is Generated Every Minute on Social Media,” We Are Soc. Media, 19.

[3] S. Wakade, C. Shekar, K. J. Liszka, and C.-C. Chan, “Text mining for sentiment analysis of Twitter data,” in Proceedings of the International Conference on Information and Knowledge Engineering (IKE), 2012, p. 1.

[4] Saul Vargas, Richard McCreadie, Craig Macdonald, and Iadh Ounis, “Comparin overall and targeted sentiments in social media during crises,” *In Tenth International AAAI Conference on Web and Social Media, 2016*.

[5] H. Liu, Y. Sun, and M. S. Kim, “Fine-grained DDoS detection scheme based on bidirectional count sketch,” *Proc. - Int. Conf. Comput. Commun. Networks, ICCCN*, 2011.

[6] Johan Bollen, Huina Mao, and Xiaojun Zeng, “Twitter mood predicts the stock market” Journal of Computational Science, 2(1):1–8, 2011.

[7] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Combining strengths, “emotions and polarities for boosting twitter sentiment analysis In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining,” page 2, *Chicago*, USA, 2013. ACM.

[8] Andreas M Kaplan and Michael Haenlein, “Users of the world unite! the challenges and opportunities of social media,” *Business horizons,* 53(1):59-68, 2010.

[9] Simon Kemp. Digital, social and mobile worldwide in 2015, special report. Available: http://wearesocial.com/uk/special-reports/digital-social-mobile-worldwide-2015.

[10] Twitter, Available: https://about.twitter.com/company.

[11] whizsky, Available: https://www.whizsky.com/2018/05/amazing-social-media-statistics-and-facts-in-2018.

[12] Matthew A Russell, “Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More,” *O'Reilly Media, Inc*, 2013.

[13] Twitter api documentation, Available: https://dev.twitter.com/overview/documentation.

[14] Klaus R Scherer, “Vocal communication of emotion: A review of research paradigms. Speech communication,” 40(1):227-256, 2003.

[15] A.Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining” In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326.

[16] Nadia FF da Silva, Eduardo R Hruschka, and Estevam R Hruschka, “Tweet sentiment analysis with classi_er ensemble,” *Decision Support Systems*, 66:170-179, 2014.

[17] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume” pages 79-86, *Association for Computational Linguistics*, 2002.

[18] Bing Liu, “Sentiment analysis and subjectivity,” *Handbook of natural language processing*, 2:627-666, 2010.

[19] Bishan Yang and Claire Cardie, “Joint inference for _ne-grained opinion extraction,” *In ACL (1)*, pages 1640-1649, 2013

[20] agdale, Rajkumar S. Vishal S. Shirsat, and Sachin N. Deshmukh, “Sentiment Analysis of Events from Twitter Using Open Source Tool,” 2016.

[21] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu, “Social media mining: an introduction,” *Cambridge University Press*, 2014.

[22] Olivier Chapelle, “Support vector machines et classification images,” 1998.

[23] letcher, T. (2009), “Support vector machines explained, tutorial paper,” Available: http://www.tristanfletcher.co.uk/SVM%20Explained.pdf.

[24] Donaldson, “Beautiful decisions: Inside bigml's decision trees,” Available: http://blog.bigml. com/2012/01/23/beautiful-decisions-inside-bigmls-decision-trees.

[25] Leo Breiman, “Random forests. Machine learning,” 45(1):5-32, 2001.

[26] Decision trees, Available: http://webtutplus.com/decision-trees/.

[27] Christopher D Manning, Prabhakar Raghavan, Hinrich Schutze, et al, “Introduction to information retrieval,” *Cambridge university press Cambridge*, 2008.

[28] Monisha Kanakaraj and Ram Mohana Reddy Guddeti, “Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques,” *In Semantic Computing (ICSC), 2015 IEEE International Conference on, pages 169-170*, IEEE, 2015.

[29] R Muhamedyev, K Yakunin, S Iskakov, S Sainova, A Abdilmanova, and Y Kuchin, “Comparative analysis of clascation algorithms,” I*n Application of Information and Communication Technologies (AICT), 2015 9th International Conference on*, pages 96-101. IEEE, 2015.

[30] Hassan Saif, Yulan He, and Harith Alani. “Semantic sentiment analysis of twitter,” *In the Semantic Web-ISWC 2012*, pages 508-524. Springer, 2012.

[31] Luiz FS Coletta, Nadia FF da Silva, and Estevam R Hruschka, “Combining class cation and clustering for tweet sentiment analysis,” *In Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 210-215. IEEE, 2014.

[32] M. A. Hearst, “Untangling text data mining,” *in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 3–10.

[33] B. Pang, L. Lee, and others, “Opinion mining and sentiment analysis,” *Found. Trends®in Inf. Retr,* (1-2), 1–135.

[34] J. Akaichi, Z. Dhouioui, and M. J. L.-H. Pérez, “Text mining facebook status updates for sentiment classification,” *in System Theory, Control and Computing (ICSTCC), 2013 17th International Conference,* 2013, 640–645.

[35] A.-H. Tan and others, “Text mining: The state of the art and the challenges,” *in Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases,* 1999, 8, 65–70.

[36] M. Rajman and R. Besançon, “Text mining: natural language techniques and text mining applications,” *in Data mining and reverse engineering*, Springer, 1998, 50–64.

[37] G. Ifrim, G. Bakir, and G. Weikum, “Fast logistic regression for text categorization with variable-length n-grams,” in Proceedings of the 14th ACM SIGKDD international.

[38] M. D. Sykora, T. W. Jackson, A. O’Brien, and S. Elayan, “National security and social media monitoring: A presentation of the emotive and related systems,” *in Intelligence and Security Informatics Conference (EISIC),* 2013 European, 2013, 172–175.

[39] M. Cheong and V. C. S. Lee, (2011). “A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter,” *Inf. Syst. Fron.,* 13(1), 45–59.

[40] K. Glass and R. Colbaugh, (2012). “Estimating the sentiment of social media content for security informatics applications” *Secur. Inform,* 1(1),3.

[41] V. D. Nguyen, B. Varghese, and A. Barker, “The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter,” *in Big Data, 2013 IEEE International Conference on,* 2013, pp. 46–54.

[42] M. M. Mostafa, (2013). “More than words: Social networks text mining for consumer brand sentiments,” *Expert Syst. Appl,* 40(10), 4241–4251.

[43] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz, (2013). "Web-scale pharmacovigilance: listening to signals from the crowd," *J. Am. Med. Informatics Assoc*, 20(3), 404–408.

[44] H. Isah, P. Trundle, and D. Neagu, "Social media analysis for product safety using text mining and sentiment analysis," *in Computational Intelligence (UKCI), 2014 14th UK Workshop on*, 2014, pp. 1–7.

[45] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, (2011). "Lexicon-based methods for sentiment analysis," *Comput. Linguist*, 37(2), 267–307.

[46] A. Z. H. Khan, M. Atique, and V. M. Thakare, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis,*" Int. J. Electron. Commun. Soft Comput. Sci. Eng*, p. 89, 2015.

[47] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *in LREc*, 2010, vol. 10, no. 2010.

[48] B. Liu, "Sentiment Analysis and Subjectivity," *Handb. Nat. Lang. Process*, vol. 2, pp. 627–666, 2010.

[49] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data," *in Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 2013, pp. 198–206.

[50] K. Gimpel et al., "Part-of-speech tagging for twitter: Annotation, features, and experiments," *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, short papers-Volume 2, 2011, pp. 42–47.

[51] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating mapreduce for multi-core and multiprocessor systems," *in High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on*, 2007, pp. 13–24.

[52] D. Gillick, A. Faria, and J. DeNero, “Mapreduce: Distributed computing for machine learning,” *Berkley, Dec*, vol. 18, 2006.

[53] S. Papadimitriou and J. Sun, “Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining,” i*n Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on, 2008,* pp. 512–521.

[54] S. Das, Y. Sismanis, K. S. Beyer, R. Gemulla, P. J. Haas, and J. McPherson, “Ricardo: integrating R and Hadoop,” *in Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010, pp. 987–998.

[55] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, “Toolkit-based high-performance data mining of large data on MapReduce clusters,” *in Data Mining Workshops, 2009. ICDMW’09. IEEE International Conference on*, 2009, pp. 296–301.

[56] A. Ghoting and E. Pednault, “Hadoop-ML: An infrastructure for the rapid implementation of parallel reusable analytics,” *in Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS09)*, 2009.

[57] E. Kouloumpis, T. Wilson, and J. D. Moore, “Twitter sentiment analysis: The good the bad and the omg!,” *Icwsm, vol. 11*, no. 538–541, p. 164, 2011.

[58] J. Lin and A. Kolcz, “Large-scale machine learning at twitter,” *in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 793–804.

[59] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of twitter data,” *in Proceedings of the workshop on languages in social media*, 2011, pp. 30–38.

[60] H. Saif, Y. He, M. Fernandez, and H. Alani, “Semantic patterns for sentiment analysis of Twitter,” *in International Semantic Web Conference*, 2014, pp. 324–340.

[61] S. Asur and B. A. Huberman, “Predicting the future with social media,” *in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 2010, (1), 492–499.

[62] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci, vol. 2*, no. 1, pp. 1–8, 2011.

[63] C. Yang, K. H.-Y. Lin, and H.-H. Chen, “Emotion classification using web blog corpora,” *in Web Intelligence, IEEE/WIC/ACM International Conference on*, 2007, pp. 275–278.

[64] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” *in Proceedings of the ACL student research workshop,* 2005, pp. 43–48.

[65] A. Go, L. Huang, and R. Bhayani, “Twitter Sentiment Analysis. Final Project Report.” *Stanford University, Department of Computer Science,* 2009.

[66] S. Ruggieri, (2002) “Efficient C4. 5 classification algorithm.,” *IEEE Trans. Knowl*. Data Eng., 14(2), 438–444.

[67] E. Kretschmann, W. Fleischmann, and R. Apweiler, (2001). “Automatic rule generation for protein annotation with the C4. 5 data mining algorithm applied on SWISS-PROT,” *Bioinformatics,* 17(10), 920–926.

[68] J. R. Quinlan and others, “Bagging, boosting, and C4. 5,” *in AAAI/IAAI,* 1, 1996, 725–730.

[69] Z. Xiaoliang, Y. Hongcan, W. Jian, and W. Shangzhuo, “Research and application of the improved algorithm C4. 5 on decision tree,” *in Test and Measurement*, 2009.

[70] L. P. Rajeswari and K. Arputharaj, (2008). “An active rule approach for network intrusion detection with enhanced C4. 5 Algorithm,” *Int. J. Commun,* Netw. Syst. Sci., 1(4), 314.

[71] M. M. Mazid, S. Ali, K. S. Tickle, and others, “Improved C4. 5 algorithm for rule based classification,” *in Proceedings of the 9th WSEAS international conference on Artificial intelligence,* knowledge engineering and data bases, 2010, pp. 296–301.

[72] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?,” *in Proceedings of the 19th international conference on World wide web,* 2010, pp. 591–600.

[73] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose.,” *in ICWSM,* 2013.

[74] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, (2015). “A survey of open source tools for machine learning with big data in the Hadoop ecosystem,” *J. Big Data,* 2(1), 24.

[75] T. Verma, R. Renu, and D. Gaur, (2014). “Tokenization and filtering process in RapidMiner,” *Int. J. Appl,* Inf. Sys, 7(2), 16–18.

[76] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature, vol.* 381, no. 6583, pp. 607–609, 1996

[77] Z. Xiaoliang, Y. Hongcan, W. Jian, and W. Shangzhuo, “Research and application of the improved algorithm C4. 5 on decision tree,” *in Test and Measurement,* 2009.

[78] Leng and P. Jiang, “A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm,” K*nowledge-Based Systems, vol.* 100, no. C, pp. 188–199, 2016.

[79] E. Kretschmann, W. Fleischmann, and R. Apweiler, (2001). “Automatic rule generation for protein annotation with the C4. 5 data mining algorithm applied on SWISS-PROT,” *Bioinformatics,* 17(10), 920–926.

[80] Ankit Kumar Soni, “Multi-Lingual Sentiment Analysis of twitter data by using classification algorithms, ” 978-1-5090-3239-6/17/$31.00©2017IEEE

[81] Min-Yuh Day, Yue-Da Lin, “Deep Learning for Sentiment Analysis on Google Play Consumer Review, ” *2017 IEEE International Conference on Information Reuse and Integration (IRI).*