ISTANBUL COMMERCE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

SENTIMENT ANALYSIS ON NEW CURRENCY IN KENYA USING TWITTER DATASET

Ibrahim Moge NOOR

Supervisor
Asst. Prof. Dr. Metin TURAN

MASTER THESIS
COMPUTER ENGINEERING DEPARTMENT
ISTANBUL – 2020

ACCEPTANCE AND APPROVAL PAGE

On 11/12/2020, **Ibrahim Moge NOOR** successfully defended the thesis, entitled **"Sentiment Analysis on New Currency in Kenya Using Twitter Dataset"** which he prepared after fulfilling the requirements specified in the associated legislation, before the jury members whose signatures are listed below. This thesis is accepted as a **Master's Thesis** by Istanbul Commerce University, Graduate School of Natural and Applied Sciences **Computer Engineering Department.**

Supervisor	**Asst. Prof. Dr. Metin TURAN** Istanbul Commerce University	…….
Jury Member	**Assoc. Prof. Dr. Serhan YARKAN** Istanbul Commerce University	……...
Jury Member	**Asst. Prof. Dr. Zeynep TURGUT** Haliç University	……..

Approval Date: 11/12/2020

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 11.12.2020 tarih ve 2020/298 numaralı Yönetim Kurulu Kararının 1. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen "Ibrahim Moge NOOR" (TC:99331353534.) adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK
Acting Head of Graduate School of Natural and Applied Sciences

ACADEMIC AND ETHICAL RULES
DECLARATION OF CONFORMITY

In this thesis study that I prepared in accordance with the thesis writing rules of Istanbul Commerce University, Institute of Science,

- I have obtained all the information and documents in the thesis within the framework of academic rules,
- I have presented all visual, auditory and written information and results in accordance with scientific ethics,
- I refer to the relevant works in accordance with scientific norms in case the works of others are used,
- I cite all the works I refer to as a source,
- I have not made any tampering with the data used,
- and that I have not submitted any part of this thesis at this university or another university as another thesis

I declare.

11/12/2020

Ibrahim Moge NOOR

CONTENTS

Page
CONTENTS i
ABSTRACT ii
ÖZET iii
ACKNOWLEDGEMENT iv
LIST OF FIGURES v
LIST OF TABLES vi
SYMBOL AND ABBREVIATIONS LIST vii
1. INTRODUCTION 1
1.1 Overview 1
1.2 Background of Demonetization Policy 1
1.3 Twitter Data 2
1.4 Classification of Sentiment Analysis 3
1.5 Objective and Limitations 4
1.6 Multilingual Tweets 5
1.7 Motivation 6
2. LITARATURE REVIEW 7
3. PROPOSED APPROACH 9
3.1 Data Collection 9
3.2 Data Set 11
3.2.1 Train data 12
3.2.2 Test data 13
3.3 Data Pre-Processing 13
3.4 Feature Extraction 15
3.4.1 The term frequency–inverse document frequency 15
3.4.2 Count vectorization 16
3.5 N-gram Model 17
3.6 Sentiment Analysis of Tweets 18
3.7 Bayesian classifier: Naive Bayes 19
3.7.1 Pros and Cons of Naive Bayes? 21
3.8 Multinomial Naïve Bayes 22
3.9 Bayesian Classifier 22
3.10 Confusion Matrix 25
3.11 Dataset Validation 26
3.11.1 Report Dataset after applied cross validation tests: 27
4. RESULT AND DISCUSSION 34
5. CONCLUTIONS AND IMPLICATIONS 38
REFERENCES 39
BIBLIOGRAPHY 42

ABSTRACT

Master Thesis

SENTIMENT ANALYSIS ON NEW CURRENCY IN KENYA USING TWITTER DATASET

Ibrahim Moge NOOR

Istanbul Commerce University
Graduate School of Sciences and Engineering
Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Metin TURAN

2020, 41 pages

Social media sites recently became popular, it is clear that it has major influence in society. Twitter is one of these sites, full of people's opinions, where one can truck sentiment express about different kinds of topics. Sentiment analysis is one of the major interesting research areas nowadays. In this work, we focused on sentimental insight into the 2019 Kenya currency replacement. Kenyans citizens expressed their reaction over new banknotes. We perform sentiment analysis of the tweets from Twitter using the Multinomial Naïve Bayes algorithm. We split our dataset using k-folder cross validation since we had limited amounts of data, so to achieve unbiased prediction of the model. We calculated unigram and bigram models and given as features to the Multinomial Naïve Bayes classifier. We found an accuracy of 70.8% when we used unigram model and 64.1% when we applied bigram model. Results show that the model reached to an acceptable accuracy of (72%) on average using unigram model.

Keywords: Machine learning, Multinomial Naïve Bayes, sentiment analysis, Twitter data.

ÖZET

Yüksek Lisans Tezi

TWİTTER VERİ KÜMESİNİ KULLANARAK KENYA'NIN YENİ PARA BİRİMİ ÜZERİNDE DUYGU ANALİZİ

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Metin TURAN

2020, 41 sayfa

Sosyal medya siteleri son zamanlarda popüler hale gelmiştir, toplumda büyük etkisi olduğu açıktır. Twitter, bu tür sitelerden biridir, insanların görüşleri ile dolu olup, farklı türlerdeki konularda duyguları ifade edebilir. Duygu analizi, günümüzde önemli ilginç araştırma alanlarından biridir. Bu çalışmada, 2019 Kenya para birimi değişimine ilişkin duygusal analize odaklandık. Kenya vatandaşları yeni banknotlar üzerindeki tepkilerini dile getirmiştir. Multinomial Naïve Bayes algoritmasını kullanarak, Twitter tweet'lerinin duygu analizini yaptık. Veri setimiz, sınırlı miktarda veriye sahip olduğundan, modelin tarafsız tahminini elde etmek için k-çapraz doğrulama yöntemi kullanarak böldük. Unigramları ve bigramlarıhesapladık ve Multinomial Naïve Bayes sınıflandırıcısına özellik olarak verdik. Unigram modelini kullandığımızda %70.8, bigram modelini uyguladığımızda %64.1 doğruluk bulduk. Sonuçlar, modelin unigram kullanarak ortalama olarak kabul edilebilir bir doğruluğa (72%) ulaştığını göstermektedir.

Anahtar Kelimeler: Duygu analizi, makine öğrenmesi, Multinomial Naïve Bayes, Twitter verileri.

ACKNOWLEDGEMENT

I would firstly like to express my deep gratitude to Asst. Prof. Dr. Metin TURAN, my thesis supervisor, for his valuable contributions for my works.

Beside my advisor, I am very grateful to my parent who always encouraged me and prayed for me throughout the time of my studies.

Ibrahim Moge NOOR
İSTANBUL, 2020

LIST OF FIGURES

Pages

Figure 1. 1.Tweet mixed with mixed with Swahili language 5
Figure 1. 2.Abbreviation used in tweets 6
Figure 3. 1.Diagrammatic representation of proposed methods 10
Figure 3. 2.Emotional distribution 12
Figure 3. 3.System flowchart 13
Figure 3. 4.Train words polarities. 17
Figure 3. 5.Sentiment analysis method 19
Figure 3. 6.Sentiment Classification Based On Emoticons 20
Figure 3. 7.Cross validation division 27
Figure 4. 1.Emotion Distribution of Data Set 34
Figure 4. 2.Emotion Distribution of Test Data Set Accuracy 35

LIST OF TABLES

Pages

Table 3. 1.Sample of raw tweets collected 11
Table 3. 2.Unigram words 18
Table 3. 3.Bigram words 18
Table 3. 4.Prediction result for test 1 using unigram 27
Table 3. 5.Prediction result for test 2 using unigram 27
Table 3. 6.Prediction result for test 3 using unigram 28
Table 3. 7.Prediction result for test 4 using unigram 28
Table 3. 8.Prediction result for test 5 using unigram 28
Table 3. 9.Prediction result for test 6 using unigram 29
Table 3. 10.Prediction result for test 7 using unigram 29
Table 3. 11.Prediction result for test 8 using unigram 29
Table 3. 12.Prediction result for test 9 using unigram 30
Table 3. 13.Prediction result for test 1 using bigram 30
Table 3. 14.Prediction result for test 2 using bigram 30
Table 3. 15.Prediction result for test 3 using bigram 31
Table 3. 16.Prediction result for test 4 using bigram 31
Table 3. 17.Prediction result for test 5 using bigram 31
Table 3. 18.Prediction result for test 6 using bigram 32
Table 3. 19.Prediction result for test 7 using bigram 32
Table 3. 20.Prediction result for test 8 using bigram 32
Table 3. 21.Prediction result for test 9 using bigram 33
Table 4. 1.Cross validation subsets 35
Table 4. 2.Unigram train data 36
Table 4. 3.Bigram train data 36

SYMBOLS AND ABBREVIATIONS LIST

CBK	Central Bank of Kenya
IDF	Inverse document frequency
MNB	Multinomial Naive Bayes
N	Negative
NB	Naïve Bayes
P	Positive
SA	Sentiment Analysis
TF	Term frequency

1. INTRODUCTION

1.1 Overview

Sentiment analysis has started long time ago and still there are a lot of researches on this topic, its common application of natural language processing where the emotion of writer are extracted from data information are distinguished whether is positive, negative or neutral (Hirst, 2012). Nowadays most people write lots of reviews, and the reviews which are available on the internet have more perfect details than the reviews from other sources (Farisi et al., 2019). In order to drive a meaningful data from people's opinions, we need to apply machine learning technique.

1.2 Background of Demonetization Policy

Demonetization is withdrawal of currency from circulation and replace the old currency with the new one (Jayati et al., 2017). The early June 2019 during the celebration of Madaraka Day, Kenya government decided to withdraw 1000 Kenya shilling note, which is equivalent to 10 dollar, from circulation by 1st October 2019. The change of old currency with new one is something started back 2010 when new constitution was promulgated, that mandated the change of currency.

The 2010 constitution commanded the Central Bank of Kenya (CBK) to spearhead the creation new notes which should be fashioned to allow the visually impaired to use them. The constitution forbids the utilization of an individual's picture on monetary standards. The old notes has images of the successor of the first president Daniel Arap Moi and first president Jomo Kenyatta, the most interesting thing is that why the Central Bank of Kenya delay for almost a decade for fulfilment of constitution. The Central Bank of Kenya claimed that its immediate decision was based on fighting corruption such illicit financial flow and money laundering, but on other hand the CBK central bank decision has face strong headwind from private sector, although is their constitutional provision command this alter of banknotes.

The Central Bank of Kenya of after its decision of new currency faced a legal suits, one of them was filed by Kenyan activists they argued that the designs of the new

generation currency notes were not subjected to public participation. These responsibility instruments are introduced on the conviction that administrators are not just political on-screen characters but on the other hand are "open authorities" who ought to work openly on the eyes of public (Akech, 2011).

1.3 Twitter Data

After demonetization in Kenya people across country post their view on demonetization on social media. Specifically we use Twitter as a source of our data set.

In the past, to find individual differences of personality in characteristic patterns of thinking, feeling and behaving was time consuming and less practical, since person had to take personality tests and answers various question but according to the a recent research, the personality of individuals can be found automatically by observing their written sentences (Park et al., 2015).

Twitter is one of the best online social network site, the microblogging service which has also become an important source of real-time events updates, and had over 48.6 million active users and 330 million active users per month, where it plays an important role in expressing our feelings (Agarwal et al., 2011), where users can share either opinions or information about product, events or politics. Each tweet is restricted to a limit of characters where user can post a short message up to 280 characters (O'Connor et al., 2011).

Tweet feature:

- Length of a tweet: the maximum characters per tweet is 280 characters, even though some user use abbreviation like 'b4','ur','u8','g8t','sry','coz','pic' which is not meaningful and grammatically correct sentence, it can be consider as sentence.
- Language: people use Twitter in various languages, though we just considered English language tweets beside Kenya being multilingual country, 90% of Twitter user tweet with English language.
- Hashtag: are formed by using the pound sign **(#)** in front of the word with no space and punctuation like #Kenyanewcurrency, it make conversation cantered

around the same topic easier to search, this help us when we extracting tweets from Twitter.

Web scraper with help of Twitter advance search was used to extract tweets data from Twitter (Hernandez et al., 2018), on related topic as discussed above.

1.4 Classification of Sentiment Analysis

The sentiment index relies critically on tracking the reference frequencies of vocabularies with positive and negative connotations (Godbole et al., 2007). The extraction of the sentiment can be in several levels, and the most common is phrase level. The others are sentence level, paragraph level and document level extraction of sentiment (McDonald et al., 2007; Korayem et al., 2012).

Sentiment analysis was considered a grouping issue. Much the same as in enormous reports, sentiments of tweets can be communicated in various manners and characterized by the presence of sentiment, i.e., if there is sentiment in the tweets, contain polar words then it is assign either positive or negative, else it is viewed as neutral. As there are words in the text of the both two classes, they don't give any significant data. The studies shows that to applied term frequency inverse document frequency (TF-IDF) metric in order to solve this kind of problem (Benhardus et al., 2013). Some authors categorize sentiment of text into six emotions sadness, anger, disgust, fear, joy and surprise (Strapparava et al., 2013).

In order to classify the sentiment behind the tweets, count the negative and positive words allocate a score for each tweet. In view of the score, the tweet will be classified into negative, positive or neutral. Extremity scores are additionally relegated to each tweet based emotional of tweets such joy, sad, happiness or anger likewise, and base on polarity such negative, positive and neutral. The supervised learning techniques need corpus of which was classified before into specifics groups so that can be used in machine learning purposes. Supervised algorithm utilizes an assign dataset where each document of training set is labelled with appropriate sentiment.These datasets are first changed over into transitional models where records are converted to vectors , so that these data can be used to feed machine

learning algorithm (Das et al., 2013). We group the sentiments of the tweets into three groups: negative, positive and neutral (Jiang et al., 2011).

However, there are some challenges in tweets sentiments:

a) Tweets post are unlike the other social media sites they are short and normally show limited sentiment signals.

b) Unbiased tweets are substantially more average than negative and positive tweets, which will as a rule be overwhelmingly positive or negative.

1.5 Objective and Limitations

The objective of this thesis is to analyze the opinion of Kenyan people on money demonetization that took place in 2019. The following steps were applied to achieve this goal.

- Data collection which is the core of our thesis work.
- Data pre-processing is cleaning or filtering of data in order to remove noise.
- The implementation of feature extraction (TF) at first part of thesis.
- The Sentiment Classification for the training dataset, find and predict the polarities of the test dataset with unigrams and bigrams as features using Multinomial Naive Bayes algorithm at the second part of thesis.
- Discuss and analyse the model result.

Sentiment analysis of multilingual tweets challenges various difficulties. Statistical methodologies require training material which is ordinarily sparse for various dialects. Then again, lexical methodologies require language explicit lexical and semantic assets. Creating these assets is very tedious and requires regularly manual work. As per our knowledge, there are chiefly two approaches that are important with regards to multilingual sentiment analysis. A corpus based approach and a dictionary based way to deal with multilingual subjectivity analysis (abstract versus objective). Inside the dictionary based methodology, an objective language subjectivity classifier is produced by interpreting a current dictionary. The corpus-based methodology constructs a subjectivity-commented on corpus for the objective

language through projection. A factual classifier is prepared on the subsequent corpus (Demirtas and Pechenizkiy, 2013).

Machine learning demands a huge dataset for training purpose in order to obtain high accuracy. The topic is new and no one collected data relating to subject before, tweets gathered and labelled manually although it is time consuming for filtering and removing the noises.

1.6 Multilingual Tweets

The utilization of Twitter as a social network in Kenya is at present developing, with this development a lot of useful data is being pass through the network. This data can give some values to researchers or scientists on the overall view of their services. However, some Kenyan Twitter users normally tweet with mixed language such as English and Swahili and this blend is normally unstructured and casual. However, most of collected data was in English language and the little remained was translated to English language.

In thesis work the collected dataset was mixed with another languages beside English, mostly Swahili language, one of the tweet mixed with Swahili words was shown, so we had to translate this word to corresponding English words. (Figure 1.1).

Figure 1.1.Tweet mixed with mixed with Swahili language

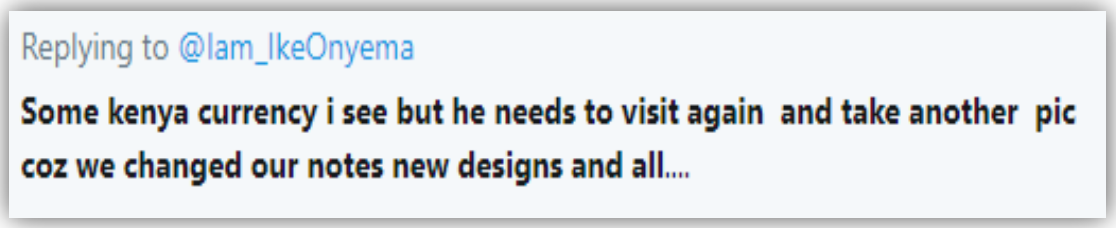

Figure 1.2.Abbreviation used in tweets

We come across abbreviation problem, the abbreviation words were changed to its original words. (Figure 1.2).

1.7 Motivation

Nowadays social media platform became popular with billions of users around world, sharing their opinions of a particular subject, as more Internet users share their opinion daily it becomes a valuable source of data. Sentiment analysis techniques are used to identify and find opinions of the authors by expressing into the polarity of positive, negative or neutral.

Obtaining data for analysis was not an easy task some years back, compare to recently where millions of internet users post their views in social media, concerning to a particular subject, one of the most popular social media sites is Twitter with daily millions of tweets. The obtained data from these sites are so important source for further analysis and decision making.

Sentiment analysis task is becoming increasingly important for various companies because of emergence of social media sites, most companies may require to track tweets of their brand to screen the impacts over time or they may want to analyse comment posted on their articles, Politian's could use to track their campaign by reviewing the comments around the internet. Sentiment analysis empowers all kinds of market research and competitive analysis.

2. LITARATURE REVIEW

Sentiment analysis is the automated process of understanding an opinion about a given subject from spoken language or written. Sentiment analysis became a trend topic of various researches and text mining has been done in past years. Multinomial Naïve Bayes method used specifically addressing recurrence in the content of the document. The Multinomial Naïve Bayes model has been introduced as an option of Naive Bayes for text classifier. In recent past years, many researchers usually regard it as the ideal Naive Bayes text classifier (Frank and Bouckaert, 2006). Multinomial Naive Bayes a family of probabilistic classifiers, the state of art of Bayesian classifier is the best since it is fast and simple text classifier (McCallum and Nigam, 1998). TF-IDF substitution relatively improved the performance of the general classifier (Susanti et al., 2017; Abbas et al., 2019). TF-IDF measures word scores effectively before characterization. TF-IDF was straightforward, actualize and process. Multinomial Naive Bayes improved considerably by applying a TFIDF change to the word features as well as weight learning (Kibriya et al., 2004). The supervise machine learning are tend to be more accurate since each of the classifiers is trained on an assortment of representative data called corpus however the supervise machine learning depends on the quality of training data as well the type of algorithm used (Chaovalit and Zhou, 2005). The collected dataset from Twitter was labelled into different polarities positive, negative and neutral, labelling data is scarce and time consuming (Srijith et al., 2013). Then was classified to their respective class using machine learning algorithms with unigrams and bigrams as features (Webb et al., 2001). The drop of accuracy in n gram for some text classification algorithm may cause by sparsity of data (Go et al., 2009).The polarity of tweets such positive and negative, neutral in tweets were studied (HaCohen-Kerner and Badash, 2016).

One of the techniques that sentiment analysis can be conducted is lexicon-based approach in which, the dictionary is made out of a lot of positive and negative assessment words, used to score the tweets either, positive, negative or neutral (Lima et al., 2015).Sentiment analysis techniques are good ways to identify and find opinions of the authors by expressing into polarity positive, negative or neutral.Hegde et.al. implemented supervised algorithms, they compared different

feature extraction determine which algorithm is best suited in term of execution time for Sentiment Analysis based on the given dataset (Hegde et al., 2015).

3. PROPOSED APPROACH

3.1 Data Collection

Demonetiztion data was collected between June and October. 1087 tweets were collected , although their was some shortage of tweets, the dataset was collected using web scraper and Twitter advance search.

The gathered data was applied an important techniques inorder to reduce the noise and dimensionality of sentence, each tweets first undergoes preprocessing step where all vague information was elimanated, then potential feature are extracted, the features are words in document, since algorithm need numerical vectors and not a textual data, in order to convert text into corresponding integers, the vectorization techniques are used. Matrix is applied to input to classification algorithm and the collected tweets was split as 89% for train data and 11% for test data. (Figure 3.1).
We follow the below five step as shown in figure 3. To analyze our tweets polarities whether is positive, negative or neutral.

Why data collection is important:

- It empowers you to find trends in the manner of individuals change their behavior and opinions overtime or in different conditions.
- It lets you portion your crowd into various client gatherings and direct unique marketing strategies at every one of the gatherings dependent on their individual needs.
- It encourages dynamic and improves the quality of decisions made.
- It helps settle issues and improve the quality of your service or products dependent on the criticism obtained.

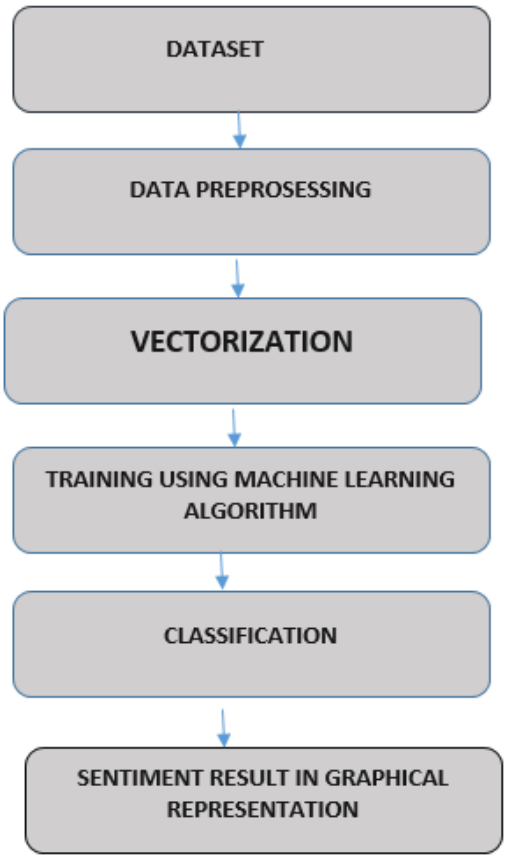

Figure 3. 1.Diagrammatic representation of proposed methods

Data Collection from Twitter: This is the essential thing of research or just said the key thing of research without information is nothing. There are numerous approaches to gathering the datasets from Twitter yet in our proposal using web Scraper. There are three distinct strides to gather the cleaned information from Twitter.

- First extracting the data from Twitter and spared in CSV record.
- Second gather the tweets text from one CSV document to spare in another CSV record.
- Third expelled duplication from tweets information.

The sample of raw data which was collected from Twitter before it was cleaned was shown. (Table 3.1).

Table 3. 1. Sample of raw tweets collected

@CBKKenya Governor @njorogep announces new Kenyan currency notes, demonetizes all older Ksh1000 banknotes in an effort to fight corruption. Kenyans have until 1st October 2019 to exchange the old notes, after which the older ones will cease to be legal tender. #MadarakaDaypic.twitter.com/Q0hKzviJA5
Tanzania Freezes Transaction of Kenyan Currency. #Tanzania #Banknotes #BankNoteNews #Banknotestreet https://buff.ly/2F2zzoW pic.twitter.com/nl0l5kFbVC
Kenya's new banknotes and the battle against corruption: A shake-up of the Kenyan currency has provoked controversy, including several court challenges. http://dlvr.it/R5xXYP pic.twitter.com/uFNQDNuMOp
The new currency notes were unveiled this month by the Kenyan central bank. The new banknotes have also sparked controversy for having the portrait of the country's founding president Jomo... https://africafeeds.com/2019/06/03/kenyan-mp-goes-to-court-over-new-bank-notes/ …
I am speaking on behalf of ODM and we are saying we accept the notes.
We advise that going forward, CBK should ensure there is nothing that can be construed to be anyone's portrait on our notes-John Mbadi
#NewCurrencyNotesKe pic.twitter.com/JwlxnbtWRc
Kenyans new currency VS Ugandan Currency. No difference at all. To shop keepers until you adapt, be careful at night. Careful! #NewCurrencyNotesKe pic.twitter.com/zscCpq533P
Have you received the new bank notes?
Our staff, Dorothy and Felix discuss the look and feel of the new currency. Remember: FEEL, LOOK, and TILT. @CBKKenya @KenyaBankers #newcurrencynoteske pic.twitter.com/sGTGi0WijP
This our Kenya so it's Matiangi is now on our #NewCurrencyNotesKe issokay pic.twitter.com/4Xb21KvEVk

3.2 Data Set

The polarity of the demonetization tweets dataset is considered for analysis. 1087 tweets, was ectracted from Twitter, between 1st June to 11th October, although their was some shortage of tweets, the dataset was collected using web Scraper and Twitter advance search.it was consist of 431 negative tweets, 332 neutral tweets and 324 positivet tweets. The collected data was split into test and train datasets: 967 tweets for training and 120 for test. Also we applied k-cross validatiom where we have assign k as 9,by spliting of dataset into 9 subsets.

As shown in pie chart above we split 88% of our dataset as training dataset and 12% test dataset. Also randomization of dataset is step we consider so that couldn't end up training some polarities only. (Figure 3. 2).

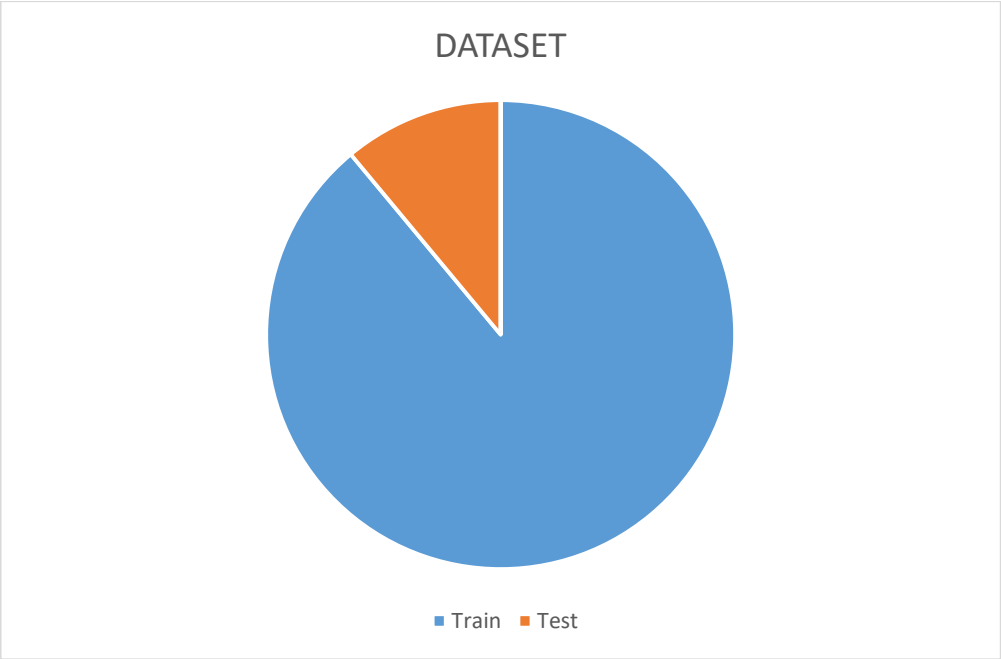

Figure 3. 2.Emotional distribution

We chose large dataset for training so that our can learn very well, the more huge data you train the better more confidence you for future prediction of your test data confront that the size of the training set relies upon the complication of the categorization problem (Hadjarian et al., 2013). However, a situation where you have small dataset is then more sophisticated and will be needed like cross validation.

3.2.1 Train data

The training data is characterized as the data that the learning algorithm uses to extricate training features, subsequently used in the classification of new data. This training data is generally included content that is recently marked as having a place with the classification classes. The utilization of the training data is to implement to build up a model, the training set is normally comprised of class marks, which distinguish to what sentiment class the set itself has a place, just as the training features themselves. Training features could incorporate unigrams, bigrams, and part of speech tags. The training features are normally gathered by the opinion class where they were extracted from. By doing this, the learning algorithm can realize which training features will be utilized to distinguish a particular opinion class.

3.2.2 Test data

The reason for testing data is to give where testing data set will be extracted as to be used by the learning algorithm to test the accuracy of the classification. The testing data is involved content that was labelled before as assign to either of the sentiment classes. From the testing data, a testing set is extricated along these lines to the extraction of the training set. The taking in algorithm utilizes features separated from the training set recently talked about above, to group contribution from the testing set, and afterward looks at its classification results, with initially labelled classification. Where the two classes are identical classification accuracy will be realise else their will error in the classification the learning algorithm has been made.

3.3 Data Pre-Processing

Data Pre-processing: is a procedure that is utilized to change over the crude information into a clean data set. The data we extracted from Twitter site was in raw format which is not feasible for the analysis. Data preprocessing is one way of preparing the data in a way that is suitable to analyse as well for decision making. (Figure 3. 3).

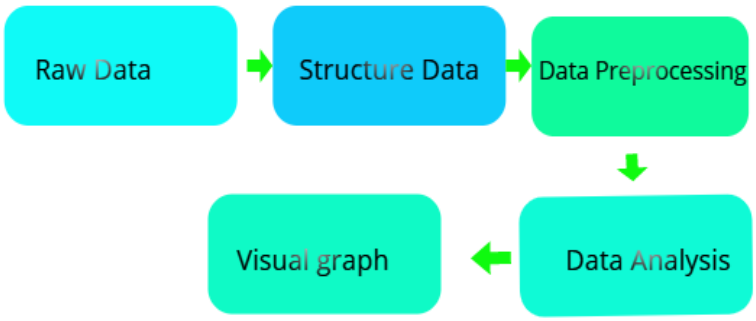

Figure 3. 3.System flowchart

Demonetiztion data was collected between june and October. then the gathered data was applied an important techniques inorder to reduce the noise and dimensionality of sentence.The data was cleaned by removing symbols, extra spaces and numbers. Also the collected tweets was mixed of hashtags '#' ,url links, annotation ' @', as shown in table 1, since it was uncessary for our decision making we elimanted, also

we remove the stop words, the stop words are common words which doesn't add a values for classification such as and, either, or to,the so on (Yu, 2008).

Stemming also was applied, we take out the root of word by triming off their end. Like Exchanged ,exchanging return to exchannge, applying stemming into our dataset give us less sparsity in our data hence more clarity. There are several stemming algorithms, we applied the most common one porter stemmer. In 1980 Martin Porter was developed the idea of Porter Stemmer Martin Porter at the University of Cambridge (Porter, 1980).

Most tweets was written in English language, but some of tweets was mixed English and Swahili language which is second national language in Kenya. as shown in figure 1, Whenever we come across that users tweets with Multilanguage's in just a single tweet we were trying to translate to English language but goodness there was no much such tweets as same time some of tweets was in abbreviation form as shown in figure 2, we were writing to its original word.

In case a raw data were used, the way the tweeters tweeted then it could take very long time and a lot energy to train the dataset, as well it could lead poor predictions and hence low accuracy.

For dataset we converted all uppercase tweets to lower case tweets, so at end we got a uniform tweets which was to train it, consider a tweets was written in upper case and another tweets in lower case both same tweets with same concept, then during trading our dataset we would consider maintain single vocabulary with upper and lower case, so my point is these would have affected the weight of vocabularies in the dataset.

Advantages of Data preprocessing

- Better evaluation and decision making.
- Speed accurate and more reliable.
- Increase productivity and lead better result

3.4 Feature Extraction

After cleaning our data, we did feature extraction. Feature selection helps in the problem of text classification to improve efficiency and accuracy (Khan et al., 2011).

Feature selection helps in the problem of text classification in order to improve accuracy and efficiency as well.

The extracted tweets was stored in unstructured format. This unstructured data supposed to be change over to a meaningful data in order to feed it to a machine learning algorithm. The algorithm need numerical vectors and not a textual data, in order to convert text into corresponding integers, the vectorization of text file to numerical vectors is done utilizing following approaches.

3.4.1 The term frequency–inverse document frequency

TF-IDF It is a numerical measurement that is aimed to reflect how important word is to corpus or a document, which is use in machine learning and text mining as weighting plan in data recovery that has additionally discovered great use in archive characterization.

When weight increases as the word frequent in document increases but is offset by the frequency of the word in document, the offset TF–IDF contains two elements term frequency and inverse document frequency, is calculate as follow:

- TF = (Frequency of a word in the document) / (Total of words in the document).
- IDF =Log ((Total number of documents) / (Number of documents containing the word)).

TF-IDF it measure relevancy of given word to a document and not frequency, if term occurs more time in the document then it has more relevance than any term in the document, a very high term frequency and very low document frequency for a given

word, the ratio of these two can give measure of the relevance of that word to the document. Actual log of the inverse document can be used instead of raw values.

TF IDF (word) = log (f + 1) × log (D /df)

3.4.2 Count vectorization

Count Vectorization gives a straightforward method to both tokenize collection of text documents and create vocabulary of known words as well encode new document by utilizing that vocabulary, which will produces a sparse representation of counts. We use Count Vectorization in our dataset as follow. We created vectors that have a dimensionality equivalent to the size of our Sentiments which is either negative, positive or neutral, so if the content data features that sentiment word, we will put a one in that dimension and rest assign zero, each time we experience that word once more, we increased the count.

Vectorization is used to accelerate the Python code without using loop. Utilizing such, a function can help in limiting the running time of code effectively. Different activities are being performed over vector, for example, dot product of vectors which is also known as scalar product as it produces single output, outer products which brings about square matrix of dimension equivalent to length X length of the vectors, Element wise multiplication which items the component of same lists and dimension of the matrix remain unchanged.

Some authors perform new features dependent on word sequences of various length from unigram, bigram till 5-gram using Naive Bayes algorithm as learning way on the feature vector, where just most importantly word with high score as per TF IDF Term frequency–inverse document frequency, were used, they demonstrated that longer sequences has no effect with the average performance (Hernández et al., 2009).

In our case only using unigram gave us better result, the type of dataset matter a lots, in our dataset there was some short tweets such as:

“When new notes ready.”
“Kenya’s new money ugly.”
“Better old currency.”

3.5 N-gram Model

Applying N-gram model in the sentiment analysis is very helpful in analysing the sentiment of document or text.in our thesis we use only unigram which refers to n-gram of size one and bigram which refers to n-gram of size two. N gram is used for improving features for supervised machine such as Naïve Bayes (Awachate and Kshirsagar, 2016).

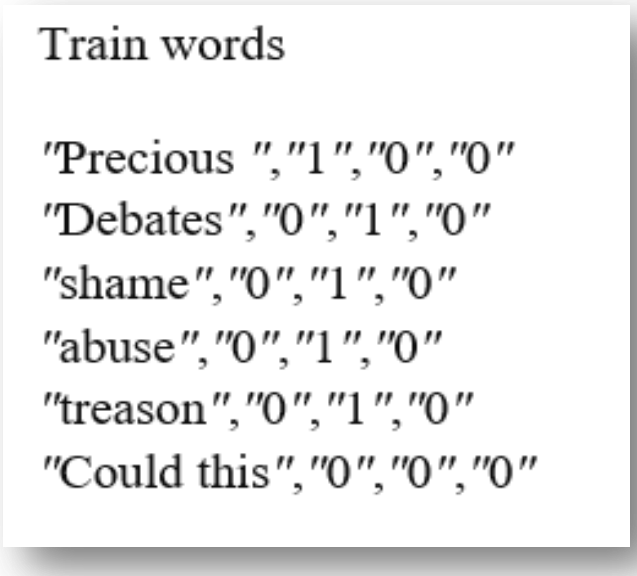

Figure 3. 4.Train words polarities.

If word is positive the first columns was assign one, else if word is negative the second columns was assign one and we assign all column zero for neutral words in case we didn’t find that word. (Figure 3. 4).

In our Dataset there was over 3000 vocabularies, most these vocabularies had low frequency so perform pruning in order to reduce over fitting and complexity of classifier also it improve our model accuracy. We have only use the most effective and information vocabularies, we train each vocabularies as its respective polarities positive, negative or neutral. We have created dataset with sentiment classification

by preparing negative word corpus, words that is disagreeing with demonetization process, positive word, and words that agreeing with demonetization events as well we created neutral words, word that neither agreeing nor disagreeing all is shown in table below. Each tweet is assessed and a numeric score is calculated. In view of this score, the labels sentiment are connected by the accompanying rules. If positive score is more than negative score was assign as positive, else if negative score is more than positive score then was assign as negative. If both negative and positive score are equal then was assign as neutral. Some of the unigram vocabularies was shown. (Table 3.2).

Table 3. 2.Unigram words

Positive words	**Negative words**	**Neutral words**
stash	crisis	launch
accessible	monopoly	visit
appreciate	difficult	release
innovate	Claims	return

Bigrams, where tokens represents two consecutive vocabularies, the most information gain and useful bigram in training model were extracted as well ignored the least ones. (Table 3.2).

Table 3. 3.Bigram words

Positive words	**Negative words**	**Neutral words**
accessible public	fake money	caution aware
advocate less	foreign currency	enough aware
agree commitment	awkward realisation	exchange ksh (Kenyan shilling)
curb fraud	flow integration	announces plan
flows counterfeit	flimsy excuses	caution public
tackle illicit	Felix discuss	laud launch
safety country	bear image	application helps
security features	court challenge	forward curb

3.6 Sentiment Analysis of Tweets

It's estimated that 80% of the world's information is unstructured and not sorted out in a pre-characterized way.

Sentiment analysis algorithms: There are various methods that can be used to implement sentiment analysis, (Figure 3.5).

Which can be group as:

1- Automatic system depends on machine learning techniques to learn the data.

2- Rule-based system which performs sentiment analysis by set of physically created principles.

3- Hybrid system combines both. Machine learning and Rule-based system approaches to address Sentiment Analysis is called Hybrid. We perform our Twitter sentiment analysis using Multinomial Naive Bayes algorithm which is a type Naive Bayes, Naive Bayes classifier to get higher accuracy and we come up a lexicon analysis which contains a words list which is negative and positive.

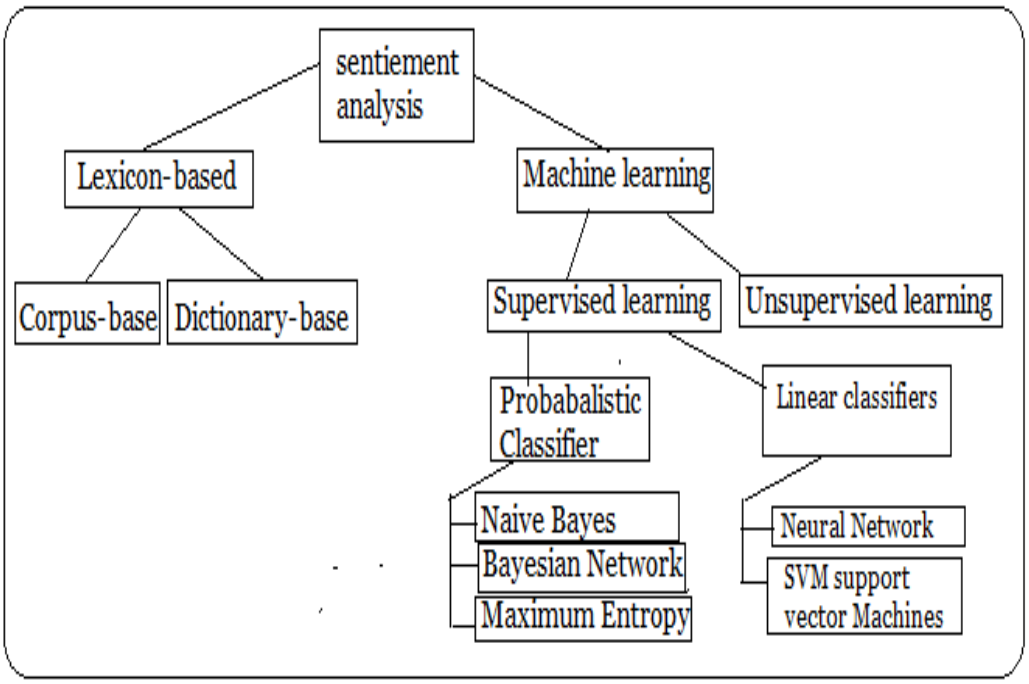

Figure 3. 5.Sentiment analysis method

3.7 Bayesian Classifier: Naive Bayes

Naive Bayes is a classification method which is based on Bayes' theorem. It is suitable for large data sets since it assumes independence between predictors and it assumes that a feature in a class which is not related to any other also is fast it only need one pass over the data. (Figure 3.6).

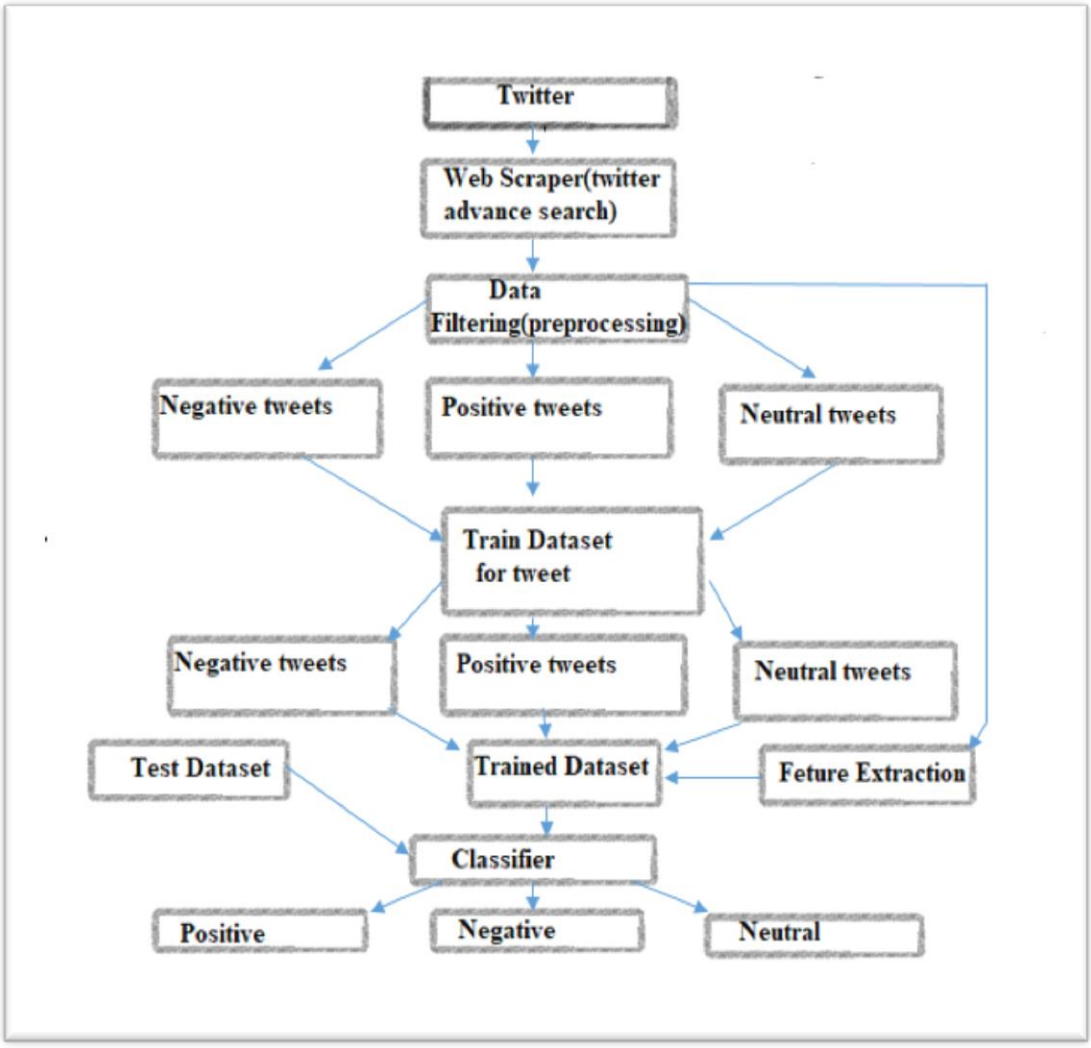

Figure 3. 6.Sentiment Classification Based On Emoticons

Naïve bayed algorithm is considered as one the best and effective in sentiment classification, it gives a better result than other classification method, to infer the tweet sentiment we use Multinomial Naïve Bayes to infer the tweets sentiment, this classifier was selected as our classifier because its simplicity and fast in sentiment classification. Naïve Bayes is a supervise classification algorithm which is based on Bayes theorem which follows probabilistic approach, where the outcome output depend on a set of independent variables that has no relation to each other, predictor variable in machine learning mode are independent to one another. The principle behind Naïve Bayes is Bayes rule which calculate the condition probability. In supervised machine learning, we have an informational index of info perceptions, each related with some right yield (a 'management signal'). The objective of the calculation is to figure out how to outline another perception to a right output.

Officially, the undertaking of supervised characterization is to take an information x and a fixed arrangement of output classes Y = y1, y2... y M and return an anticipated class ∈ Y. For text grouping, we'll here and there talk about c (for "class") rather than y as our yield variable, and d (for "report") rather than x as our info variable. In the supervised circumstance we have a preparation set of N reports that have each been hand-named with a class: (d1, c1)... (dN, cN). We will likely get familiar with a classifier that is equipped for mapping from another report d to its right class c ∈C. A probabilistic classifier moreover will disclose to us the likelihood of the perception being in the class. This full dissemination over the classes can be helpful data for downstream choices; abstaining from settling on discrete choices at an early stage can be valuable when consolidating frameworks. Numerous sorts of machine learning calculations are utilized to manufacture classifiers. This section presents naive Bayes; the accompanying one presents strategic relapse. These embody two different ways of doing characterization. Generative classifiers like naive Bayes fabricate a model of how a class could create some information. Given a perception, they return the class well on the way to have created the perception. Discriminative classifiers like calculated relapse rather take in what highlights from the information are generally helpful to separate between the various potential classes. While discriminative frameworks are frequently increasingly precise and henceforth more generally utilized.

3.7.1 Pros and Cons of Naive Bayes?

Pros of Naive Bayes:

- It is simple and quick to predict class of test data. It additionally perform well in multi class predict.
- When assumption of autonomy holds, a Naive Bayes classifier performs better compare with different other machine learning classifiers.
- It perform well if there should be an occurrence of straight out information factors contrasted with numerical variable(s). For numerical variable, ordinary dispersion is expected (ringer bend, which is a solid presumption).

Cons:

- If categorical factor has a class in test data, which was not seen in training dataset, at that point model will dole out a 0 (zero) likelihood and will be not able to make a forecast. This is regularly known as "Zero Frequency". To explain this, we can utilize the smoothing strategy. One of the least complex smoothing methods is called Laplace estimation.
- On the opposite side Naive Bayes is otherwise called an awful estimator, so the likelihood yields from predict probability are not to be paid attention to as well.

3.8 Multinomial Naïve Bayes

Multinomial Naïve Bayes which is a type of Naïve Bayes. Multinomial Naïve Bayes method used to represent the recurrence in the text of the document and improve accuracy than simply checking for the word occurrence. It is a probabilistic classifier, and there are two fundamental methodologies you could take, to train our model in order to recognize the polarity tweets (positive, negative or neutral).A supervised which is an object of interest in this thesis. The first of all we collected would ask from you to gather labelled data, and train the algorithm, in a supervised way how each word in a grouping relates to the result of in general sentence being positive, negative or neutral. This methodology requires physically marked data, which is regularly tedious, and not constantly conceivable. Unsupervised learning is that you don't give any past presumptions and definitions to the model about the result of factors you feed into it, you just supplement the information and need the model to become familiar with the structure of the data itself.

3.9 Bayesian Classifier

Bayesian classifier, Bayesian classifier introduced by Thomas Bayes in 1763 It is based on the theorem. Bayesian classifiers statistical classification techniques by the researchers due to its speed and its performance in calculation. It is a frequently preferred algorithm. The events to be classified independently from each other this theorem predicts which class the data belongs to. It does not require a starting time before classification and all that will be it processes entire data sets for

classifications. Easy applicability, most situations Good results and high performance are among the advantages of Bayes' theorem. Countable. However, since variables are dependent on each other in practice, there is a problem in modelling the relationship (Rennie et al., 2003). This theorem runs between conditional probabilities and marginal probabilities of random variables. Shows that there is a relationship. P (A) is the first probability of A, and P (B) is the probability of B. P (A \ B) shows the contingent probability. Equation mathematically Bayes' theorem.

Naive Bayes, from Bayes' theorem in classification of text created used, understandable and easily applicable It is one of the simplest machine learning algorithms. With this method, the class of the target attribute of an instance there are possibilities to belong to the value.

According to basian model parameters are random and data is fixed since we train model, Bayesian methods work for arbitrary number of data, basian model train data by computing the posterior of the probability of the parameters given datum by using Bayes formula, computing the posterior distribution, for the classification that is probability of parameters given the training dataset using Bayes formulas can be computed the prediction of the probability of the test data given training dataset.

In the sentiment detector model we consider trio polarities (x,) comprises of label y ∈ (1,0,0), (0,1,0) and(0 ,0,0) positive, negative and neutral respectively is a V dimensional vector of word count for a vocabulary of size V, thus x(j) is frequency of the vocabulary it appear in the tweets sentences, we have to characterize dissemination on x as its model which can likewise be considered as a joint likelihood circulation of all values in x.

P (x|θ) ⇔ p(x (1), . . . , x (v) |θ)

The document feature vectors in the multinomial document model, catches the frequency of the words. Consider the below example to get it clear about the concept: N_i.

Let x_i alone the multinomial model element vector for the i th documentD_i. The t th component of x_i written X_{it}, is the count of the occasions word w_t happens in documentD_i.

Let $n_i = \sum_t x_{it}$ be the all-out number of words in report D_i.
Let P(w_t|C) again be the likelihood of word wt happening in class C, this time estimated utilizing the word recurrence data from the archive include vectors.

The probability of each word happening in the document is totally independent other words occurrences. We would then be able to compose the Document probability P (D_i |C) as a multinomial conveyance, where the quantity of attracts compares to the length of the archive, and the extent of drawing thing t is the likelihood of word type t happening in a record of class C,

P (w_i |C).

P (D_i |C) ~P (x_i |C) = $\frac{n_i !}{\prod_{t=1}^{|v|} x_{it}!} \prod_{t=1}^{|v|}$ P (w_i |C)$^{x_{it}}$

$$\alpha \quad \prod_{t=1}^{|v|} \text{P } (w_i \text{ |C})^{x_{it}}$$

We have labelled our train dataset L classes (Negative, positive and neutral), we can estimated multinomial sentiment classification as follow:
- Characterize the Vocabulary; the quantity of words in the vocabulary characterizes the element of the feature vectors.
-We consider the training dataset counts

M the all-out number of reports,, M_L the quantity of reports marked with class C=L, for each class L=1, . . . , L,
• x_{it} The recurrence of word wt in report Di, figured for each word wt in V.

Then we get probability P (w_i | C=l) as the prior P(C=L).

We found first the prior probability of our document by just dividing number of document of that class (either positive, negative and neutral) by total number of document.

$$P(c) = \frac{Nc}{N}$$

We calculate for word given in a class, $P(w|c) = \frac{Count(W,C) + 1}{Count(c) + |V|}$,the Addison of 1 and v (vocabulary), for smoothing purpose in case some word got zero count. We have used the sum of logs to avoid underflow.

$$\Pr(c) \propto \prod_{w=1}^{|v|} \Pr(w|c)^{fw}.$$

We have used the sum of logs to avoid underflow.

$$\Pr(c)\ \alpha\ \log(\pi c \Pi \Pr(w|c)\ fw|V|w=1).$$

3.10 Confusion Matrix

We use a confusion matrix to summarize the performance and prediction results of our classification algorithm, Confusion matrix it give us a better idea of our classification model.

The dataset utilized for the experiments was divided into 3 classes, positive, negative and neutral. For a given classifier and a record there are six possible results: true positive, false positive, true negative and false negative, true neutral, false neutral. In the event that the tweet was labelled positive and is classified positive it is considered true positive, else if is classified as either negative or neutral then it considered as false positive.

If a tweet was labelled as negative and is classified negative it is considered true negative, else if is classified as either positive or neutral then it considered as false negative, similarly to neutral.

3.11 Dataset Validation

Validation is a significant step that permits us to test the accuracy of our model. The most well-known ways to deal with validation are

- Hold out technique
- Cross validation strategy.

Validation is a significant step that permits us to test the accuracy of our model. The most well-known ways to deal with validation are hold out technique and cross validation strategy. In the hold out strategy, part of the information is held out for testing and the rest of the dataset are utilized for training the classifier.

The cross validation technique, by comparison, we split our dataset into testing and training, the information is checked a few times and every division or part of the training dataset is get the opportunity to be utilized in the training as well as testing stages.

When recorded our first result we applied the k-cross validation in order to be sure strength of train model. In cross validation strategy, the dataset was divided into 9 divisions. One is utilized for testing and 8 for training in the primary run. In the subsequent run, an alternate part is utilized for testing and 8 parts for training except the initially test data. The runs proceed until each part or division is allowed to be part of the training dataset and the testing data. (Figure 3.7).

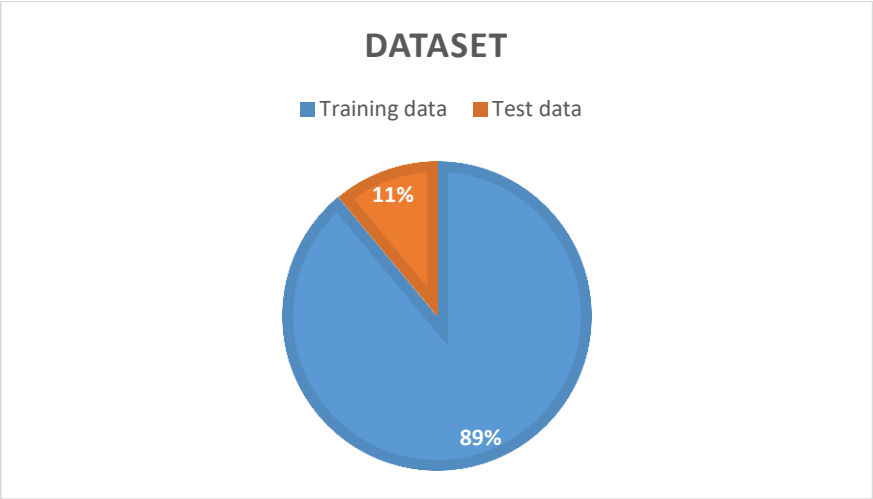

Figure 3. 7.Cross validation division

3.11.1 Report Dataset after applied cross validation tests:

Unigram results

Test 1

Table 3. 4.Prediction result for test 1 using unigram

	Precision	recall	f1-score	support
1	0.80	0.81	0.80	43
2	0.71	0.63	0.61	43
3	0.61	0.68	0.64	34
micro avg	0.71	0.71	0.71	120
macro avg	0.71	0.71	0.71	120
weighted avg	0.71	0.71	0.71	120

Test 2

Table 3. 5.Prediction result for test 2 using unigram

	Precision	recall	f1-score	support
1	0.33	0.50	0.40	6
2	0.97	0.89	0.92	96
3	0.54	0.68	0.60	19
micro avg	0.83	0.83	0.83	121
macro avg	0.61	0.69	0.64	121
weighted avg	0.87	0.83	0.85	121

Test 3

Table 3. 6.Prediction result for test 3 using unigram

	Precision	recall	f1-score	support
1	0.41	0.64	0.50	14
2	0.87	0.77	0.82	81
3	0.39	0.42	0.41	26
micro avg	0.68	0.68	0.68	121
macro avg	0.56	0.61	0.57	121
weighted avg	0.72	0.68	0.69	121

Test 4

Table 3. 7.Prediction result for test 4 using unigram

	Precision	recall	f1-score	support
1	0.78	0.76	0.77	37
2	0.79	0.74	0.76	61
3	0.48	0.59	0.53	22
micro avg	0.72	0.72	0.72	120
macro avg	0.68	0.70	0.69	120
weighted avg	0.73	0.72	0.72	120

Test 5

Table 3. 8.Prediction result for test 5 using unigram

	Precision	recall	f1-score	support
1	0.72	0.83	0.77	65
2	0.50	0.32	0.39	19
3	0.53	0.49	0.51	37
micro avg	0.64	0.64	0.64	121
macro avg	0.58	0.54	0.56	121
weighted avg	0.63	0.64	0.63	121

Test 6

Table 3. 9.Prediction result for test 6 using unigram

	Precision	recall	f1-score	support
1	0.77	0.81	0.79	67
2	0.82	0.61	0.70	23
3	0.53	0.58	0.55	31
micro avg	0.71	0.71	0.71	121
macro avg	0.71	0.67	0.68	121
weighted avg	0.72	0.71	0.71	121

Test 7

Table 3. 10.Prediction result for test 7 using unigram

	Precision	recall	f1-score	support
1	0.77	0.81	0.79	67
2	0.82	0.61	0.70	23
3	0.53	0.58	0.55	31
micro avg	0.71	0.71	0.71	121
macro avg	0.71	0.67	0.68	121
weighted avg	0.72	0.71	0.71	121

Test 8

Table 3. 11.Prediction result for test 8 using unigram

	Precision	recall	f1-score	support
1	0.60	0.75	0.67	24
2	0.81	0.66	0.72	32
3	0.85	0.85	0.85	65
micro avg	0.78	0.78	0.78	121
macro avg	0.75	0.75	0.75	121
weighted avg	0.79	0.78	0.78	121

Test 9

Table 3. 12.Prediction result for test 9 using unigram

	Precision	recall	f1-score	support
1	0.67	0.76	0.72	38
2	0.92	0.69	0.79	49
3	0.56	0.68	0.61	34
micro avg	0.71	0.71	0.71	121
macro avg	0.72	0.71	0.71	121
weighted avg	0.74	0.71	0.72	121

Bigram result

Test 1

Table 3. 13.Prediction result for test 1 using bigram

	Precision	recall	f1-score	support
1	0.78	0.65	0.71	43
2	0.60	0.86	0.70	43
3	0.55	0.35	0.43	34
micro avg	0.64	0.64	0.64	120
macro avg	0.64	0.62	0..61	120
weighted avg	0.65	0.64	0.63	120

Test 2

Table 3. 14.Prediction result for test 2 using bigram

	Precision	recall	f1-score	support
1	0.23	0.50	0.32	6
2	0.88	0.79	0.84	96
3	0.32	0.37	0.37	19
micro avg	0.71	0.71	0.71	121
macro avg	0.48	0.55	0.50	121
weighted avg	0.76	0.71	0.73	121

Test 3

Table 3. 15.Prediction result for test 3 using bigram

	Precision	recall	f1-score	support
1	0.46	0.86	0.60	14
2	0.87	0.88	0.87	81
3	0.69	0.35	0.46	26
micro avg	0.76	0.76	0.76	121
macro avg	0.67	0.69	0.64	121
weighted avg	0.78	0.76	0.75	121

Test 4

Table 3. 16.Prediction result for test 4 using bigram

	Precision	recall	f1-score	support
1	0.84	0.70	0.76	37
2	0.72	0.79	0.75	61
3	0.27	0.27	0.27	22
micro avg	0.67	0.67	0.67	120
macro avg	0.61	0.59	0.60	120
weighted avg	0.67	0.67	0.67	120

Test 5

Table 3. 17.Prediction result for test 5 using bigram

	Precision	recall	f1-score	support
1	0.79	0.85	0.81	65
2	0.48	0.68	0.57	19
3	0.62	0.41	0.49	37
micro avg	0.69	0.69	0.69	121
macro avg	0.63	0.65	0.62	121
weighted avg	0.69	0.69	0.68	121

Test 6

Table 3. 18.Prediction result for test 6 using bigram

	Precision	recall	f1-score	support
1	0.78	0.78	0.78	67
2	0.51	0.83	0.63	23
3	0.47	0.26	0.33	31
micro avg	0.65	0.65	0.65	121
macro avg	0.59	0.62	0.58	121
weighted avg	0.65	0.65	0.64	121

Test 7

Table 3. 19.Prediction result for test 7 using bigram

	Precision	recall	f1-score	support
1	0.47	0.83	0.60	30
2	0.30	0.78	0.64	27
3	0.83	0.38	0.52	64
micro avg	0.58	0.58	0.58	121
macro avg	0.61	0.66	0.58	121
weighted avg	0.67	0.58	0.56	121

Test 8

Table 3. 20.Prediction result for test 8 using bigram

	Precision	recall	f1-score	support
1	0.44	0.71	0.54	24
2	0.57	0.72	0.64	32
3	0.69	0.45	0.54	65
micro avg	0.57	0.57	0.57	121
macro avg	0.57	0.62	0.57	121
weighted avg	0.61	0.57	0.57	121

Test 9

Table 3. 21.Prediction result for test 9 using bigram

	Precision	recall	f1-score	support
1	0.57	0.68	0.62	38
2	0.65	0.76	0.70	49
3	0.56	0.29	0.38	34
micro avg	0.60	0.60	0.60	121
macro avg	0.59	0.58	0.257	121
weighted avg	0.60	0.60	0.59	121

4. RESULT AND DISCUSSION

We have 1087 tweets from Twitter, between June to October We extracted tweets from Twitter using web scrapper with help of Twitter advance search It seems that negative tweets was little bit higher compare to positive and neutral tweets. Analyses was done on this marked datasets utilizing the term frequency–Inverse document frequency (TF–IDF) extraction procedure. We use the framework where the pre-processor is applied to the raw sentences which make it increasingly fitting to comprehend. The dataset collected was label to their respective polarities, positive, negative and neutral. (Figure 4.1).

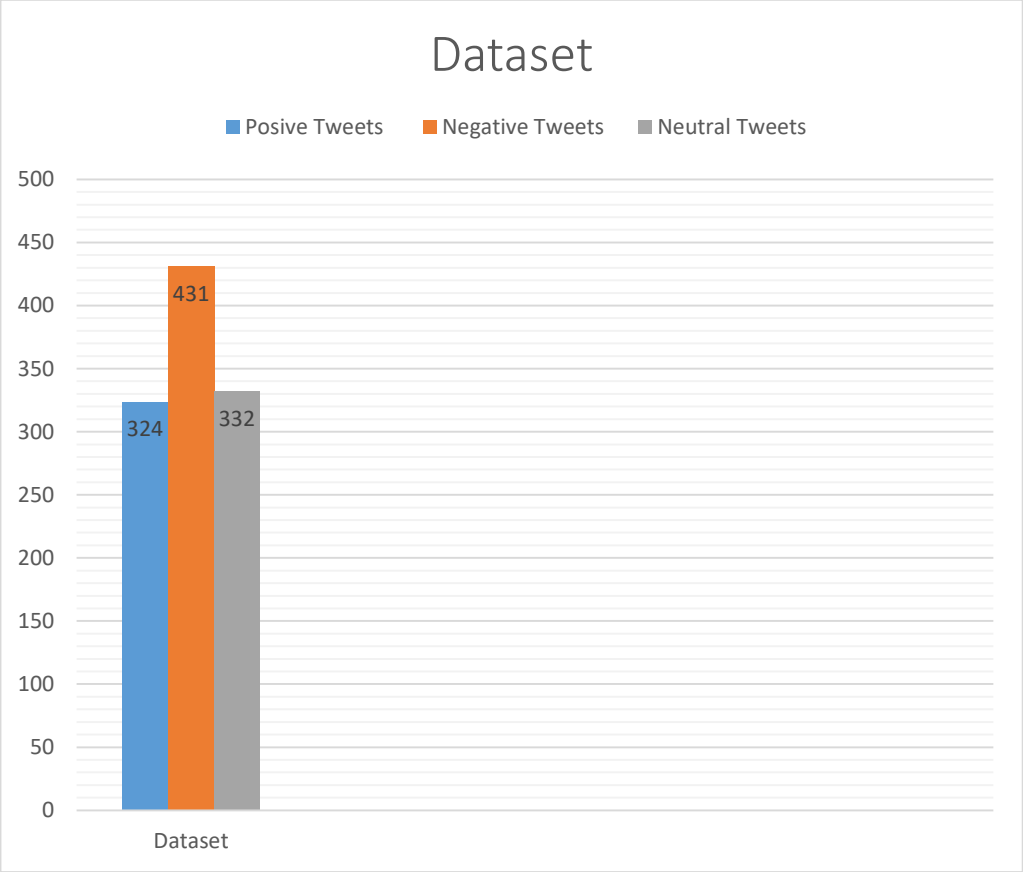

Figure 4. 1.Emotion Distribution of Data Set

Some tweets that we have collected was short, it was not possible to apply higher n-gram, since when n-gram length increases, and the number of time you will perceive any given n-gram will diminish. The drop of accuracy in bigram may cause sparsity. The more sparse data is, the more terrible you can train it. Thus, regardless of that a higher-request n-gram model, the more data in our context will contain and more will lead to over-fitting, a situation where your training data will memorizes instead of

learning which will cause poor prediction, to avoid these situation we prefer to use only lower n-gram model.

We applied cross validation we divided almost 9 equal subsets, in order to reduce bias. We train the dataset on a subset and utilize the other subset to assess the model's performance. To decrease fluctuation we achieve various rounds of cross-validation with various subsets from the same dataset. (Table 4.1).

Table 4. 1.Cross validation subsets

Train Data	967	966	966	967	966	966	966	966	966
Test Data	120	121	121	120	121	121	121	121	121

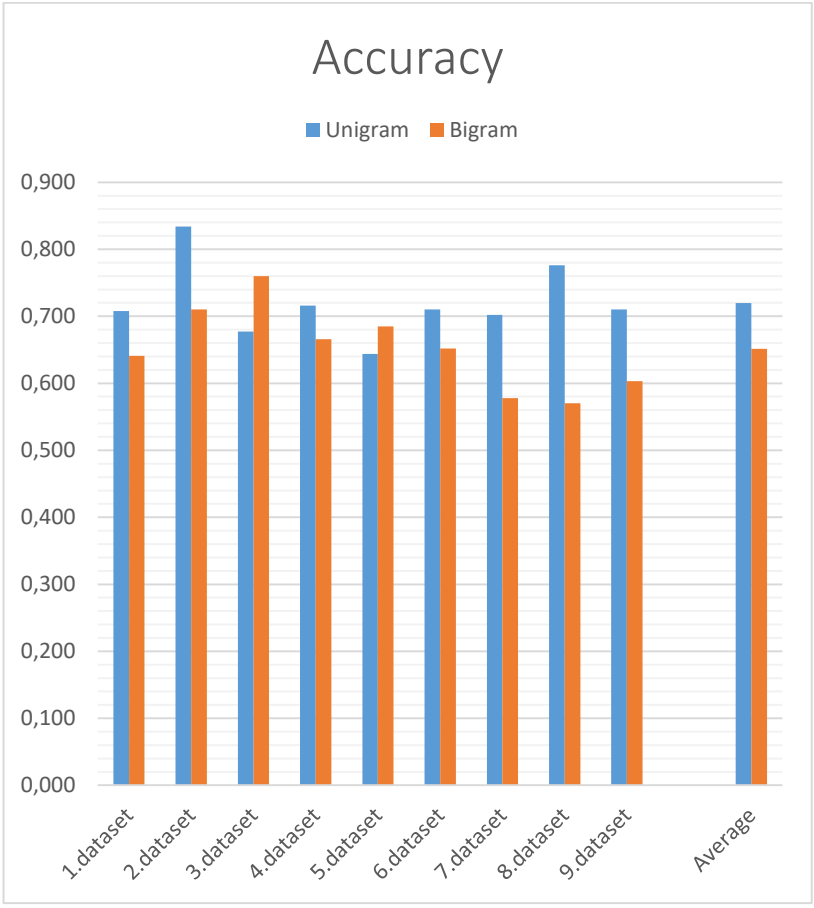

Figure 4. 2.Emotion Distribution of Test Data Set Accuracy

1. Dataset is the test data, untrained dataset we obtained 70.8% of accuracy when used unigram compared when we used bigram 64.1% accuracy, 2 to 9 dataset is train dataset. (Fıgure 4.2).

Our vocabulary was rich but most of our vocabulary wasn't have enough frequency with this reason our unigram perform better that our bigram model as shown figure 9 above.in most of time bigram perform better than unigram but in our case we were working with limit dataset since demonetization took place in short period we couldn't maintained to collect a huge data.

The overview of accuracy, recall and Precision of the dataset is as shown below. When used unigram was 71% were obtained and when bigram were used 65% accuracy were obtained. (Table 4.2), (Table 4.3).

Table 4. 2.Unigram train data

	Precision	recall	f1-score	support
1	0.69	0.77	0.73	323
2	0.83	0.73	0.77	431
3	0.62	0.65	0.64	332
micro avg	0.72	0.72	0.72	1086
macro avg	0.71	0.72	0.71	1086
weighted avg	0.73	0.72	0.72	1086

Table 4. 3.Bigram train data

	Precision	recall	f1-score	support
1	0.64	0.75	0.69	323
2	0.69	0.80	0.74	431
3	0.57	0.36	0.44	332
micro avg	0.65	0.65	0.65	1086
macro avg	0.64	0.64	0.63	1086
weighted avg	0.64	0.65	0.64	1086

The general output shows the unigram outperform better than when used bigram, for all result precision and f1-score. The prediction of neutral, except the recall of negative it gave us better result when applied bigram. Neutral prediction was perfect

when we used a unigram compare when we used bigram where most of the time it assumed either polarities positive and negative.

A sum of 120 tweets were saved aside for exactness testing As clarified above table 4.4, two kinds of feature extraction unigram, bigram were compared.

According to the output given it shows that unigram gives the best accuracy of 70.8% when applied to a Multinomial Naive Bayesian classifier. The accuracy of classification could be improved when training the dataset is increased. Twitter sentiment analysis is not an easy task since there is difficult in distinguishing sentiment words that carrying the polarity from the Tweets to solve this issues data pre-processing was applied then features extracted was done.

5. CONCLUTIONS AND IMPLICATIONS

In this thesis work we perform Twitter sentiment analysis to understand people's opinions on demonetization. The gathered dataset which was extracted from Twitter using web scrapper. The data size was limited we had to work on small size and to avoid bias prediction we have applied cross validation. Multinomial Naive Bayes (MNB) algorithms are implemented as well as unigram and bigram as our feature. Analyses was done on this marked datasets utilizing the term frequency–Inverse document frequency (TF–IDF) extraction procedure. We use the framework where the pre-processor is applied to the raw sentences which make it increasingly fitting to comprehend and after test data was executed, we compare both unigram and bigram. Unigram feature extraction were highest with 70.8% for test data compare to bigram which score an accuracy of 64%.

The accuracy is very acceptable when looked at the fact that we used only around 967 training dataset to train. This proceeds to show that if the number of Tweets in training dataset is expanded the precision will undoubtedly raise.

REFERENCES

Abbas, M., Memon, K.A., Jamali, A.A., Memon, S., Ahmed, A., 2019. Multinomial Naive Bayes Classification Model for Sentiment Analysis, 19(3), 62- 67.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., 2011. Sentiment Analysis of Twitter Data, in Proceedings of the Workshop on Language in Social Media, 30–38.

Akech, M., 2011. Abuse of Power and Corruption in Kenya: Will the New Constitution Enhance Government Accountability, Indiana Journal of Global Legal Studies, 18(1), 341-394.

Awachate, P. B., Kshirsagar, V.P., 2016. Improved Twitter Sentiment Analysis Using N Gram Feature Selection and Combinations, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 5(9), 151-160.

Benhardus, J., Kalita, J., 2013. Streaming Trend Detection in Twitter, Int. J. Web Based Communities, 9(1), 122-138.

Chaovalit, P., Zhou, L., 2005. Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches, 38th Hawaii International Conference on System Sciences, 1-9.

Das, B., Chakraborty, S., 2018. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation, IEEE, 1-6.

Demirtas, E., Pechenizkiy, M., 2013. Cross-lingual Polarity Detection with Machine Translation. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining 1-8.

Farisi, A.A., Sibaroni, Y., Al Faraby, S., 2019. Sentiment Analysis on Hotel Reviews Using Multinomial Naïve Bayes Classifier. Journal of Physics: Conference Series, 1192, 1-10.

Frank, E., Bouckaert, R.R., 2006. Naive Bayes for Text Classification with Unbalance. Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, 503-510.

Go, A., Bhayani, R., Huang, L., 2009. Twitter Sentiment Classification Using Distant Supervision, 1(12), 6.

Godbole, N., Srinivasaiah, M., Skiena, S., 2007. Large-Scale Sentiment Analysis for News and Blogs, Google Inc, 1-4.

HaCohen-Kerner, Y., Badash, H., 2016. Positive and Negative Sentiment Words in a Blog Corpus Written in Hebrew, Procedia Computer Science, 96, 733–743.

Hadjarian, A., Zhang, J., Cheng, S., 2013. An Empirical Analysis of the Training and Feature Set Size in Text Categorization for e-Discovery, Senior Associate Deloitte Financial Advisory, 1-7.

Hegde, B., Nagashree, H., Prakash, M., 2018. Sentiment Analysis of Twitter data: A Machine Learning Approach to Analyse Demonetization Tweets, 5(6), 1874-1880.

Hernandez, G., Arnulfo, R., Ledeneva, Y., 2009. Word Sequence Models for Single Text Summarization, International Conferences on Advances in Computer-Human Interactions, 44-48.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., Perez-Meana, H., 2018. A Web Scraping Methodology for Bypassing Twitter API Restrictions, 1-7.

Hirst, G., 2012. Sentiment Analysis and Opinion Mining. Morgan. Claypool. 143p, California.

Jayati, G., Chandrasekhar, C.P., Patnaik, P., 2017. Demonetisation Decoded A Critique of India's Currency Experiment, Routledge, 99p, New york.

Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T., 2011. Target-dependent Twitter Sentiment Classification, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 151-160.

Khan, A., Bahurdin, B.B., Khan, K., 2011. An Overview of E-Documents Classification, International Conference on Machine Learning and Computing, 544-552.

Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G., 2004. Multinomial Naive Bayes for Text Categorization Revisited, Advances in Artificial Intelligence, 3339, 488–499.

Korayem, M., Crandall, D., Abdul-Mageed, M., 2012. Subjectivity and Sentiment Analysis of Arabic: A Survey, Advanced Machine Learning Technologies and Applications, 322, 128–139.

Lima, A.C, E.S., De Castro, L.N., Corchado, J.M., 2015. A POLARITY ANALYSIS FRAMEWork for Twitter Messages, Applied Mathematics and Computation, 270, 56–767.

McCallum, A., Nigam, K., 1998. A Comparison of Event Models for Naive Bayes Text Classification, in In Aaai-98 Workshop on Learning for Text Categorization, 41–48.

McDonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J., 2007. Structured Models for Fine-to-Coarse Sentiment Analysis, in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 432–439.

O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A., 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, presented at the Fourth International AAAI Conference on Weblogs and Social Media, 122-129.

Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.P., 2015. Automatic Personality Assessment Through Social Media Language, Journal of Personality and Social Psychology, 1-18.

Porter, M.F., 1980. An Algorithm for Suffix Stripping, 14(3), 130-137.

Rennie, J.D., Shih, L., Teevan, J., Karger, D.R., 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In Proceedings of the 20th International Conference on Machine Learning, 616-623.

Srijith, P.K., Shevade, S., Sundararajan, S., 2013. Semi-supervised Gaussian Process Ordinal Regression, in Advanced Information Systems Engineering, 7908, 144–159.

Strapparava, C., Mihalcea,R., 2007. SemEval-2007 Task 14: Affective Text, in Proceedings of the Fourth International Workshop on Semantic Evaluations, 70–74.

Susanti, A.R., Djatna, T., Kusuma, W.A., 2017. Twitter's Sentiment Analysis on Gsm Services using Multinomial Naïve Bayes, TELKOMNIKA, 15(3), 1354-1361.

Webb, G.I., Pazzani, M.J., Billsus, D., 2001. Machine Learning for User Modeling, Kluwer Academic, 19-29.

Yu,B., 2008. An Evaluation of Text Classification Methods for LITERARY STUDY, Literary and Linguistic Computing, 23(3), 327–343.

BIBLIOGRAPHY

Name Surname : Ibrahim Moge NOOR

Birth Place and Year : Kenya, 20/03/1993

Marital Status : Single

Foreign Languages : English, Turkish, Arabic, Swahili and Somali

E-posta : engineermoge@gmail.com

Education Status

High school : Young Muslims, 2007-2010

Undergraduate : Karabuk University, Engineering, Computer Engineer, 2013-2017.

Post graduate : İstanbul Commerce University, Institute of Science, Department of Computer Engineer, 2018-2020.

Professional Experience

IETT Internship 2015

Arge Bilişim Internship 2016

Publications

Noor, I.M., Turan, M., 2020. Sentiment Analysis on New Currency in Kenya using Twitter Dataset. International Journal on Informatic for Development, 8(2), 81-87.