Utilizing Hashtags for Sentiment Analysis of Tweets using Machine Learning

By

Rafia Jahan

153-35-1362

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

Fall – 2019

APPROVAL

This thesis titled on "Utilizing Hashtags for Sentiment Analysis of Tweets using Machine Learning", submitted by **Rafia Jahan**, **153-35-1362** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan
Professor and Head
Department of Software Engineering
Faculty of Science and Information Technology — **Chairman**
Daffodil International University

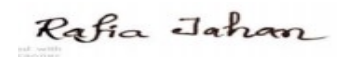
Dr. Md. Asraf Ali
Associate Professor — **Internal Examiner 1**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Asif Khan Shakir
Lecturer — **Internal Examiner 2**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Dr. Md. Nasim Akhtar
Professor — **External Examiner**
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

DECLARATION

It hereby declares that this thesis has been done by me under the supervision of Mr. Md. Shohel Arman, Lecturer, Department of Software Engineering, Daffodil International University. It is also declared that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.



Rafia Jahan

Student ID: 153-35-1362

Batch**:** 18th

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

Certified by:

Mr. Md. Shohel Arman

Lecturer

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENT

Above all else, I am appreciative to the All-powerful God that he gives ME the ability to finish a definitive proposition.

I would wish to explicit my exceptional thanks of feeling to my educator Mr. Md. Shohel Arman Joined Countries office gave ME the brilliant opportunity to attempt to this great proposition regarding the matter Using hashtags for estimation investigation of tweets exploitation AI, that furthermore helped ME in doing heaps of examination and that I came to get a handle on concerning such a major measure of new things.

I would wish to explicit my outrageous true feeling and thankfulness to all or any of our teachers of programming bundle Designing division for their sort encourage, liberal suggestion and backing all through the examination.

I moreover explicit my inclination to all or any of my friend's, senior, junior who, straightforwardly or in a roundabout way, have Lententide their guide during this endeavor.

Last anyway not the littlest sum, I may wish to pass on my family for natural procedure to ME at the essential spot and supporting ME profoundly for a mind-blowing duration.

Rafia Jahan
153-35-1362

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

LIST OF NOMENCLATURE

Terms	Nomenclature
SVM	Support Vector Machine
NLP	Natural Language Processing
BOW	Bag of Words
TF-IDF	Term Frequency-Inverse Document Frequency
NLTK	Natural Language Toolkit
POS	Parts-of-Speech
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
WUP	Wu-Palmer Similarity
SLR	Systematic Literature Review

ABSTRACT

Sentiment Analysis could be a strategy for programmed extraction of data from the assessment of others in regards to some unequivocal subject or downside. the idea of Conclusion mining and Assumption Examination apparatus is to "process an assortment of indexed lists for a given thing, creating a stock of item traits (quality, choices and so forth.) and collecting assessment". anyway with the entry of your time extra eye catching applications and advancements appeared during this space and at present its principle objective is to frame PC ready to recognize and produce feelings like human. This paper can endeavor to represent considerable authority in the crucial meanings of Supposition Mining, investigation of phonetic assets required for Feeling Mining, barely any AI methods on the possibility of their use and significance for the examination, investigation of Notion characterizations and its various applications.

In this paper associate degree way to deal with feeling investigation is presented that utilizations twitter dataset that is the preeminent standard microblogging stage for the assignment of estimation examination. Using the hashtags from twitter dataset we tend to fabricate an estimation classifier that may affirm unbiased, positive and negative conclusion for a report. Test examination show this arranged systems region unit prudent and performs higher than past arranged procedures.

CHAPTER 1: INTRODUCTION

1.1 Problem outline

Sentiment analysis is a term that alludes to the use of semantic correspondence process, content examination, and etymology in order to set up the edge of a speaker or creator toward a specific point. Essentially, it works out whether a book is communicating notions that square measure positive, negative, or impartial. Notion examination is a sublime gratitude to find anyway people, fundamentally clients, feel a few unequivocal theme, item, or thought.

The starting point of slant examination are regularly determined to the Nineteen Fifties, when conclusion investigation was fundamentally utilized on composed paper records. Today, be that as it may, notion investigation is wide wont to mine abstract data from content on the net, just as writings, tweets, web journals, online networking, news stories, audits, and remarks. this should be possible utilizing a kind of totally various procedures, just as IP, insights, and AI techniques.

So on unsettle the higher than challenge, we tend to style our system by removing important expression which gives higher tf-idf (Salton and McGill, 1983) worth for given expressions and managing WordNet (Mill operator et al.,1995).

1.2 Motivation

During search of information investigation subjects and related theory papers on this I went over here. An expanding Number of individuals are changing their perspective by following online networking like twitter. The most contemporary subjects are featured on

twitter by utilizing hashtag. Also, most NLP methods depend on AI to get significance from human dialects.

In this way I understand on the off chance that I can find the region of positive and negative idea about a specific point it is simpler to take next choice.

1.2.1 Elements of our study

We studied "Natural language processing" for preprocess of given text datasets and topics over the documents. WordNet for context collection and selection process of accurate phrase.

1.2.2 Interest behind our research

Because of the readers who does not want to read full documents to know whether a topic is negative or positive to the people worldwide, sentiment analysis can help them to know about it easily. What people are expecting from a specific topic it will be easily known.

1.3 Research Questions

RQ1: What are the sentiment analysis algorithms existing in the current-state-of-art?

RQ2: How to figure out the most covering models for sentiment analysis?

1.4 Research Objectives

- ➢ To present the sentiment analysis algorithms existing in the current-state-of-art.
- ➢ To figure out the most covering model for sentiment analysis.

CHAPTER 2: BACKGROUND

2.1 Sentiment Analysis

Sentiment analysis alludes to the administration of estimations, feelings, and abstract content. The feeling might be sketched out as a read of or point of view toward a situation or occasion. It ordinarily includes numerous sorts of liberal emotions: bliss, delicacy, trouble, or longing. One could plot the estimation names as an extremity (e.g., positive, impartial, and negative) or numerous sorts of passionate inclination (e.g., furious, upbeat, dismal, pleased). The meaning of supposition names can affect the consequence of the slant examination, along these lines we need to plot the feeling marks thoroughly. It intends to comprehend the viewpoint or assessment of a creator as to an exact point or objective.

The point of view may repeat his/her sentiment and investigation, his/her compelling situation. The interest of supposition investigation is raised gratitude to build request of breaking down and organizing concealed data that originates from the web-based social networking inside the sort of unstructured information.

For the most part, assessment investigation plan to locate enthusiastic extremity of content - in favored case - if content is certain, negative or nonpartisan.

There square measure to a great extent two types of feeling investigation techniques:

- lexicon-based
- machine learning

Lexicon-based: Dictionary based for the most part ways store delineated rundown of positive and negative words, with their valence. rule activity content to search out all far-popular words, and just consolidates their individual outcomes. There might be a few expansions, to incorporate some linguistic standards, similar to nullification or assumption modifier.

in order to perform conclusion examination by exploitation dictionary based for the most part approach, an opinion vocabulary should be made.

These strategies square measure upheld call trees like k-Closest Neighbors (k-NN), Restrictive Arbitrary Field (CRF), Covered up Andrei Markov Model (Gee), Single Dimensional Classification(SDC),and requested negligible advancement (SMO), related with procedures of assessment characterization.

Machine Learning Based: ML ways need explained information sets - rundown of writings with conclusion physically perceived. In the event that we tend to take enormous amount of such messages, and feed some AI rule (like neural systems or SVM) with them, it'll become familiar with the best approach to recognize slant precisely. underneath this framework, there square measure 2 sets, especially a preparation set and a test set.

Ordinarily the information set that is gathered from totally various sources and whose conduct and yield esteems square measure far-celebrated to US falls into the class of instructing informational indexes. In differentiation with this, the information sets whose qualities or conduct square measure obscure to US square quantify known as investigate informational indexes. AI approach has 3 classifications:

i) supervised;

ii) semi supervised; and

iii) unsupervised .

This methodology is equipped for mechanization and may deal with enormous amount of data subsequently these square measure horrendously suitable for assessment investigation. in this manner during this examination i'm going for AI approach for higher result.

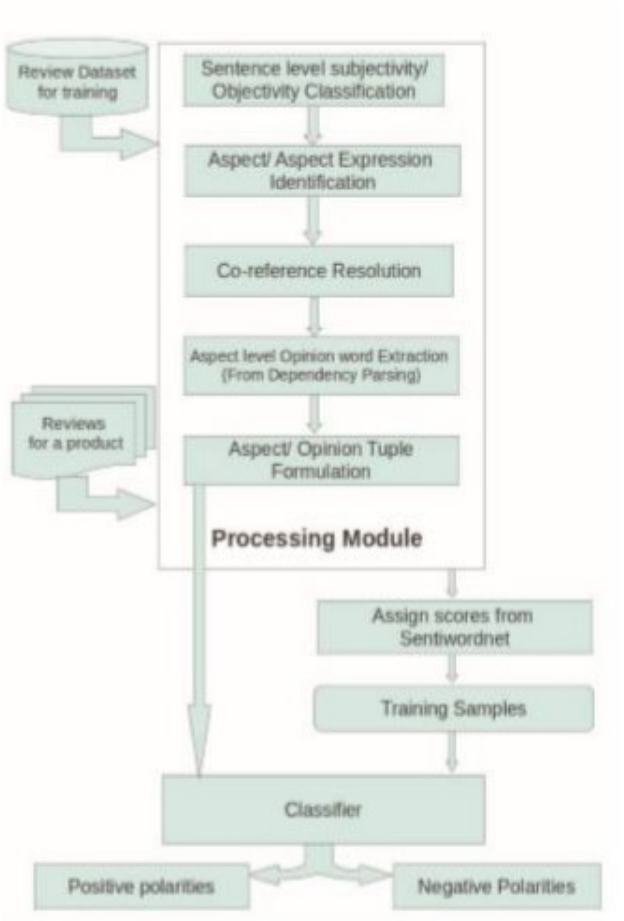


Figure 2. 1: Sentiment Analysis

2.2 Sentiment Analysis Algorithms

There are several algorithms for sentiment analysis. The top most popular ones include SVM , Naïve Bayes, Maximum Entropy Classifier.

2.2.1 Support Vector Machine (SVM)

The first SVM algorithmic program was created by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik prescribed how to frame nonlinear classifiers by applying the Part stunt to most extreme edge hyperplanes. the present standard manifestation (delicate edge) was arranged by Corinna Cortez and Vapnik in 1993 and imprinted in 1995.

A Support Vector Machine (SVM) is a regulated AI algorithmic program which will be utilized for every characterization and relapse capacities. SVMs are a ton of normally utilized in order issues and perse, this is frequently what we'll focus during this post.

SVM works alright on littler cleaner dataset. It offers a great deal of precision than elective calculations. Furthermore, conjointly It is a ton of affordable because of it utilizes a set of training focuses.

SVM is responsible for finding the decision limit to isolate entirely unexpected classes and expand the edge. along these lines it works pleasant for conclusion examination as civilian army has entirely unexpected classification like positive, negative and impartial.

2.2.2 Naïve mathematician Classifier (NB)

A naïve mathematician classifier could be a clear probabilistic classifier bolstered applying hypothesis with tough (innocent) freedom assumptions.Bayes' hypothesis was named once the Reverend Bayes (1702–61), UN office considered an approach to figure a conveyance for the possibility parameter of a measurable dispersion

Kinds of Naïve Bayes Classifier:

Multinomial Naive Bayes:

This is to a great extent utilized for record arrangement drawback, i.e whether a report has a place with the class of sports, governmental issues, innovation and so forth. The highlights/indicators used by the classifier ar the recurrence of the words blessing inside the record.

Bernoulli Naïve Bayes:

This is much the same as the multinomial credulous mathematician anyway the indicators ar Boolean factors. The parameters that we tend to use to anticipate the class variable take up exclusively values agreed or no, for instance if a word occurs inside the content or not.

Gaussian Naive Bayes:

At the point when the indicators take up a constant worth and aren't unmistakable, we tend to expect that these qualities ar inspected from an ordinary dissemination.

2.2.3 Maximum Entropy Classifier (ME):

The guideline of Maine was starting elucidated by E. T. Jaynes in 2 papers in 1957 any place he focused on a characteristic correspondence between regular way of thinking and information hypothesis. uncommonly, Jaynes offered a substitution and truly broad guideline why the Gibbsian strategy of normal way of thinking works. He contended that the entropy of normal way of thinking and furthermore the information entropy of information hypothesis ar fundamentally steady issue. Therefore, characteristic way of thinking should be seen even as a particular utilization of a general instrument of coherent thinking and information hypothesis.

2.2.4 Logistic Regression Classifier (LR) :

Strategic relapse is considered a summed up direct model because of the outcome ceaselessly relies upon the include of the information sources and parameters. Or on the other hand in

elective words, the yield can't depend upon the product (or remainder, and so on.) of its parameters! The objective of supply relapse is to appropriately anticipate the class of result for singular cases exploitation the first miserly model. To achieve this objective, a model is framed that highlights all indicator factors that ar supportive in anticipating the reaction variable. supply relapse is another system obtained by AI from the area of measurements. it's the go-to strategy for double classification varification (issues with 2 class esteems).

2.3 Reason for Using SVM in Our Research

There territory a few calculations to use for conclusion examination. with the exception of my dataset SVM shows the best outcome. the most rule of Support vector machine is to see straight separators inside the pursuit house which may best separate the different classes. As I investigation on content information SVM is one among the best choices. because of content information zone unit obviously fit to SVM attributable to the appropriated idea of content. SVM is one among the principal practical and right equation than the contrary grouping calculations.

As a rule terms Support Vector Machines zone unit horrendously goodworking once you have Brobdingnagian assortment of shifted alternatives. For content characterization during a sack of words model it works higher. SVMs will with proficiency play out a non-straight order exploitation what's known as the piece stunt, verifiably mapping their contributions to highdimensional component zones.

There are so many good terms of SVM. a number of them are-

- The horrendously idea of the arched improvement strategy guarantees secure optimality. the appropriate response will undoubtedly be an overall least and not a local least.

• SVM is Partner in Nursing recipe that is proper for each directly and nonlinearly distinguishable information (utilizing bit stunt). the sole issue to attempt to will be to returned up with the regularization term, C.

• SVMs function admirably on small likewise as high dimensional information zones. It works viably for high-dimensional datasets attributable to the very reality that the nature of the training dataset in SVM is typically described by the amount of help vectors rather than the spatial property. but all extraordinary instructing models territory unit evacuated and furthermore the training is intermittent, we'll get indistinguishable ideal isolating hyperplane.

• SVMs will work successfully on littler instructing knowledgesets as they don't have certainty the entire information.

2.4 Uses of SVM

SVM is a directed AI calculation which can be utilized for order or relapse issues. It utilizes a strategy called the bit stunt to change your information and afterward dependent on these changes it finds an ideal limit between the potential yields.

The goal of the help vector machine calculation is to discover the hyperplane that has the most extreme edge in a N-dimensional space(N — the quantity of highlights) that unmistakably characterizes the information focuses.

Information focuses falling on either side of the hyperplane can be ascribed to various classes. Additionally, the component of the hyperplane relies on the quantity of highlights. On the off

chance that the quantity of info highlights is 2, at that point the hyperplane is only a line. In the event that the quantity of information highlights is 3, at that point the hyperplane turns into a two-dimensional plane.

Support vectors are information indicates that are nearer the hyperplane and impact the position and direction of the hyperplane. Utilizing these help vectors, we amplify the edge of the classifier. Erasing the help vectors will change the situation of the hyperplane. These are the focuses that assist us with building our SVM.

2.4.1 Three impressive research using Support Vector Machine in recent years in our aspect

Table 2. 1: Five Impressive Work Using SVM

No.	Author	Used Model	Years	Problem Domain
1	Nurulhuda zainuddin et al..,2014	SVM	2014	Sentiment analysis
2	Raisa Varghese et al.,2013	SVM	2013	Aspect Based Sentiment Analysis
3	Abd. Samad Hasan Basariaet al.,2012	SVM	2012	Movie review

2.4.1.1 Research 1

Sentiment Analysis using Support Vector Machine (2014)

In their examination, the dataset was utilized to prepare a supposition order model dependent on Support Vector Machine(SVM). Here the calculation utilizes n-grams and distinctive weighting plan as a contribution to the classifier.From the perceptions, it tends to be reasoned that unigrams beat other n-grams models for both datasets while Paired Occurences (BO)and

TFIDF weighting plan assumes a significant job in removing highlights that are the most old style for Taboada Corpus just as Ache Corpus. Subsequently the paper shows that by utilizing highlight determination of chi-square will significantly advance the classification consequence of exactness for both datasets.

2.4.1.2 Research 2

Aspect Based Sentiment Analysis using Support Vector Machine Classifier (2013)

The assessment investigation issue is an exploration zone in which numerous difficult issues are to be tended to. In the work done here, the assignment of finding the viewpoints or highlights of a specific item is done naturally. With sufficient managed preparing, solid outcomes with a normal precision pace of 77.98% could be accomplished. The pre-preparing steps are relatively difficult, due to the unstructured idea of regular language content. This is a major issue to be considered and practically powerful arrangements can be given to tackle it. In our work, we utilized the coreference resolver, reliance parsing and Sentiwordnet together to create the planned outcomes.

All in all, the client audits may contain proclamations that have suppositions about viewpoints eventhough the perspective name may not be available in the sentence. Sentences with mockeries may likewise show up. Such sentences need a second request translation by the peruser separated from the simple significance passed on by the words in the sentence. Such cases happen in survey writings, and it is incredibly difficult for a machine to separate data from them. So in the present work done, such cases are not considered.

2.4.1.3 Research 3

Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization (2012)

This examination has demonstrated that PSO influence the exactness of SVM after the hybridization of SVM-PSO. The best precision level that gives in this examination is 77% and has been accomplished by SVM-PSO after information purifying. Then again, the exactness

level of SVM-PSO still can be improved utilizing upgrades of SVM that may be utilizing another mix or variety of SVM with other advancement technique.

This examination likewise has accomplished for opinion grouping utilizing SVM with a correlation among n-grams (unigram, bigram and trigram) strategy and highlight weighting (TF and TF-IDF). In future work the utilization of more blend of ngrams and highlight weighting that gives a superior precision level than this investigation will be considered. The work done in this exploration is just identified with grouping assessment into two of the classes (paired order) that is a positive class and negative class. Later on improvement, a multiclass of slant order, for example, positive, negative, nonpartisan, etc may be thought about.

CHAPTER 3: SYSTEMATIC MAPPING STUDY

3.1 Search Strategy

The exploration strategy tends to the pursuit string we express, the extent of the examination, the time of search, the hunt technique, and the consideration/rejection criteria we place. Pertinent has a wide inclusion of the best in class, the pursuit string ought to be cautiously Decided.

3.1.1 Search String

I shaped the inquiry string following to the first objective and the exploration question what I have set. The strings ought to be light, in order to acquire different outcomes and spread the subjects correctly. I have utilized the OR Boolean administrators to interface the principal terms and their counterparts. A definitive inquiry string is:

("Sentiment Analysis" OR "Sentiment Orientation" OR "Opinion mining") AND ("Twitter" OR "Social Media") AND ("Hashtag" OR "Hashtagging")

In order to use the string formed with the AND/OR Boolean operators, I have used the accessible advanced search in all database.

CHAPTER 4: RESEARCH EXPERIMENT

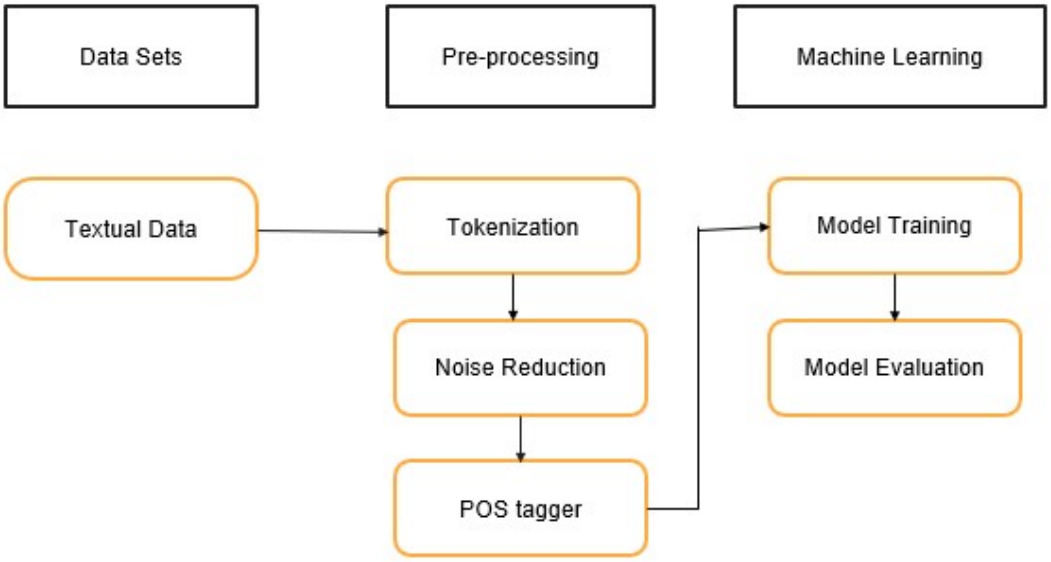


Figure 4 1: Proposed architecture

4.1 Text Preprocessing

Quantitative goals requests that we adjust our reports inside numerical information. Permitting the explanation that word arrangement might be affected of with least expenses for thought (see Grimmer and Stewart, 2013, for exchange) and a 'sack of words' depiction utilized. Research is ordinary practice (some subset of) any progressively double preprocessing (Denny et al., 2018) moves in developing the fitting report term grid. We presently clarify these in scarcely any subtleties, since certain are the focal point of our examination work.

4.1.1 Tokenizing Content

Tokenization is the demonstration of separating an arrangement of strings into pieces, for example, words, catchphrases, expressions, images and different components called tokens. Tokens can be singular words, states or even entire sentences. During the time spent tokenization, a few characters like accentuation marks are disposed of. Now and then division is utilized to move to the order of a huge lump of content inside pieces more prominent than words, while tokenization is held for the examination procedure which results especially in words.

Tokenize process

```
In [12]: tokenized_tweet = combi['tidy_tweet'].apply(lambda x: x.split())
```

4.1.2 Stop words Expelling

After complete tokenization, we have to evacuate stop words as in light of the fact that we needn't bother with all words by any means.

Stop words (Denny et al., 2018) are mean a lot of regularly utilized words in any language. When assembling the list, most motors are modified to expel certain words from any list section. The rundown of words that are not to be included is known as a stop list. Stop words are considered unimportant for looking through purposes since they happen much of the time in the language for which the ordering motor has been tuned. So as to spare both reality, these words are dropped at ordering time and afterward overlooked at search time. For instance, in the event that we search any question as "how to create slant examination with hashtag", If the internet searcher chooses to discover site pages that got the articles "how", "to" "create", "supposition", "investigation", "with", "hashtag" the web index is attempting to get a great

deal of additional pages that contain the articles "how", "to", "with" than pages that get data around subject displaying and naming in light of the fact that the articles "how", "to" and "with" are so consistently utilized in the English language. In this way, on the off chance that we overlook these two articles, the internet searcher can truly concentrate on theme demonstrating pages that get the watchwords: "create" "assessment" "examination" "hashtag" – which would all the more almost raise pages that are really of intrigue. This is only the fundamental worry for motivation stop words.

4.1.3 Stemming words

Stemming process

```
In [13]: from nltk.stem.porter import *
         stemmer = PorterStemmer()

         # apply stemmer for tokenized_tweet
         tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for i in x])
```

Stemming is the way toward diminishing intonation in words to their root structures, for example, mapping a gathering of words to a similar stem regardless of whether the stem itself is definitely not a legitimate word in the Language.. For instance, if the client questions the plural custom of a word (works, working, worked), the internet searcher recognizes and return fitting substance that relates the particular type of the root word (work).

4.2 Data Visualizing with WordCloud

Word Cloud is an information representation procedure utilized for speaking to content information in which the size of each word shows its recurrence or significance. Noteworthy printed information focuses can be featured utilizing a word cloud. Word mists (otherwise

called content mists or label mists) work in a basic way: the more a particular word shows up in a wellspring of printed information, (for example, a discourse, blog entry, or database), the greater and bolder it shows up in the word cloud.

A word cloud is an assortment, or group, of words delineated in various sizes. The greater and bolder the word shows up, the more regularly it's referenced inside a given book and the more significant it is.

```
In [16]: # create text from all tweets
         all_words = ' '.join([text for text in combi['tidy_tweet']])

         from wordcloud import WordCloud
         wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(all_words)

         plt.figure(figsize=(10, 7))
         plt.imshow(wordcloud, interpolation="bilinear")
         plt.axis('off')
         plt.show()
```

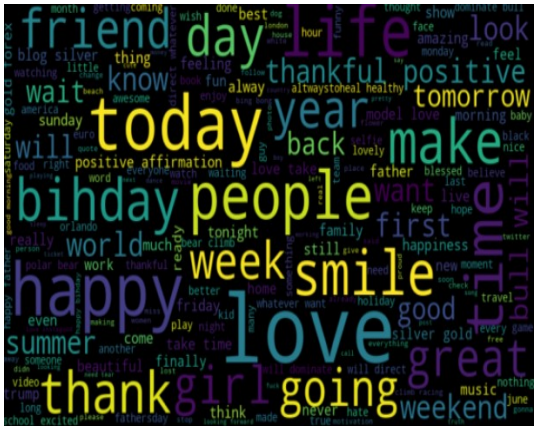


Figure 4.2: WordCloud Data

4.3 Extracting Hashtag Data

A hashtag is a one of a kind procedure for sorting messages. With regards to web-based social networking, the hashtag is utilized to draw consideration, to compose, and to advance. Hashtags got their beginning in Twitter as a method for making it simpler for individuals to

discover, pursue, and add to a discussion. Presently a-days feeling investigation has become a weapon to discover individuals' conclusion in a specific theme. As in this paper I was chipping away at twitter information, hashtags are significant device to discover assessment of individuals. Considering hashtag consistently improves opinion grouping.

```
In [19]: # function to collect hashtags
         def hashtag_extract(x):
             hashtags = []
             for i in x:
                 ht = re.findall(r'#(\w+)', i)
                 hashtags.append(ht)
             return hashtags
```

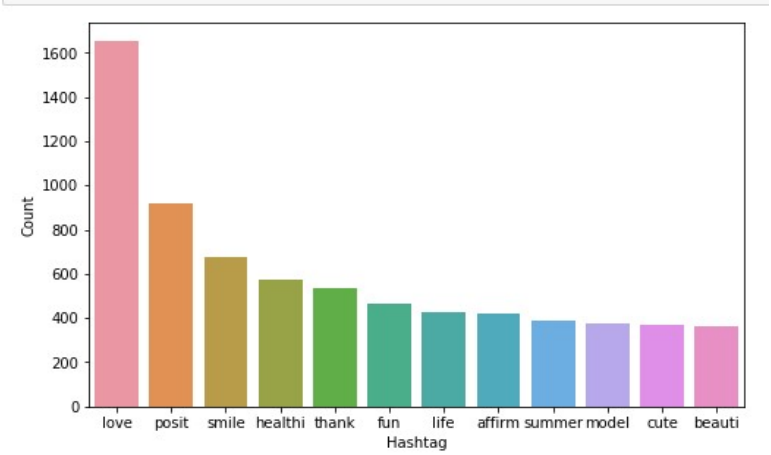


Figure 4.3: Positive data

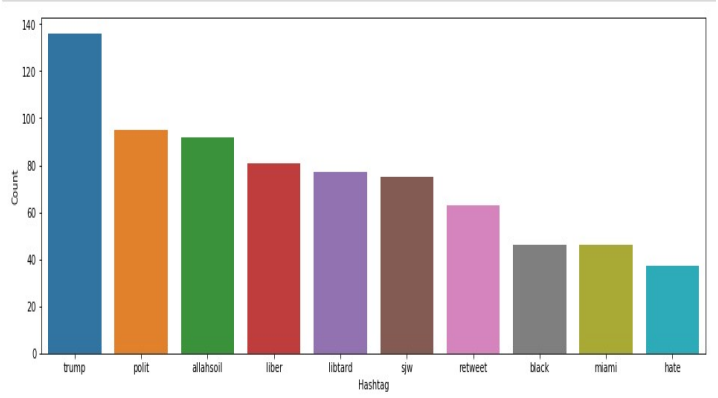


Figure 4.4: negative data

4.4 Extracting Features from Cleaned Tweets

4.4.1 Bag of Words (BOW):

The pack of-words model is a method for speaking to content information when displaying content with AI calculations. It is easy to comprehend and execute and has seen extraordinary accomplishment in issues, for example, language displaying and record order.

A pack of-words is a portrayal of content that depicts the event of words inside a record. It includes two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

For instance, if there are three sentence in an archive like " I am upbeat", " I am the best" and "that was the period of stupidity" at that point BOW will plan the content as "I", "am", "glad", "the", "best", "that", "was", "age", "of", "silliness" . after that it will make record vector for each line independently. In the event that you need to see the vector for first sentence the vector will resemble [1,1,1,0,0,0,0,0,0,0]

Furthermore, the subsequent archive will resemble [1,1,0,0,1,0,0,0,0,0]

```
In [23]: from sklearn.feature_extraction.text import CountVectorizer

         bow_vectorizer = CountVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')

         # bag-of-words feature matrix

         bow = bow_vectorizer.fit_transform(combi['tidy_tweet'])

In [24]: len(bow_vectorizer.get_feature_names())

Out[24]: 1000
```

4.4.2 TF-IDF:

TF (Term Frequency) gauges the recurrence of a word in a record.

TF = (Number of time the word happens in the content) / (Complete number of words in content)

IDF: IDF (Inverse Document Frequency) quantifies the position of the particular word for its importance inside the content. Stop words which contain superfluous data, for example, "an", "into" "and" convey less significance despite their event.

IDF = (All out number of reports/Number of records with word t in it)

In this manner, the TF-IDF is the result of TF and IDF:

TF-IDF = TF * IDF

```
In [25]: from sklearn.feature_extraction.text import TfidfVectorizer
         tfidf_vectorizer = TfidfVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')

         # TF-IDF feature matrix
         tfidf = tfidf_vectorizer.fit_transform(combi['tidy_tweet'])
```

4.3 Training SVM Model

I have partitioned the information into preparing and testing sets. Right now is an ideal opportunity to prepare our SVM on the preparation information. Scikit-Learn contains the svm library, which contains worked in classes for various SVM calculations. Since we will play out a characterization task, we will utilize the help vector classifier class, which is

composed as SVC in the Scikit-Learn's svm library. This class takes one parameter, which is the portion type. This is significant. On account of a straightforward SVM we just set this parameter as "direct" since basic SVMs can just characterize directly distinguishable information. We will see non-direct portions in the following area.

The fit strategy for SVC class is called to prepare the calculation on the preparation information, which is passed as a parameter to the fit technique. Execute the accompanying code to prepare the calculation:

As clarified in Dumais and Chen (2000) and Ache et al (2002). Given a classification set, C = {+1, −1} and two pre-classified preparing sets, i.e., a positive example set, Tr+ = ∑ = n I 1 (di , +1) and a negative example set, Tr - =∑ = n I 1 (di , −1), the SVM finds a hyperplane that isolates the two sets with greatest edge (or the biggest conceivable good ways from the two sets), as outlined in Fig. 1. At pre-handling step, each preparation test is changed over into a genuine vector, xi that comprises of a lot of significant highlights speaking to the related archive, di. Thus, Tr + =∑ = n I 1 (xi, +1) for the positive example set and Tr − =∑ = n I 1 (xi, −1) for the negative example set. In such manner, for ci =+1, w · xi + b > 0, and for ci = - 1, w · xi + b < 0. Thus, T +, T − {ci · (w · xi + b) ≥ 1} turns into an advancement issue characterized as pursues:

limit (1/2) ||w||2 , subject to ci · (w · xi + b) ≥ 1. The outcome is a hyperplane that has the biggest separation to xi from the two sides. The classification assignment would then be able to be planned as finding which side of the hyperplane a test falls into.

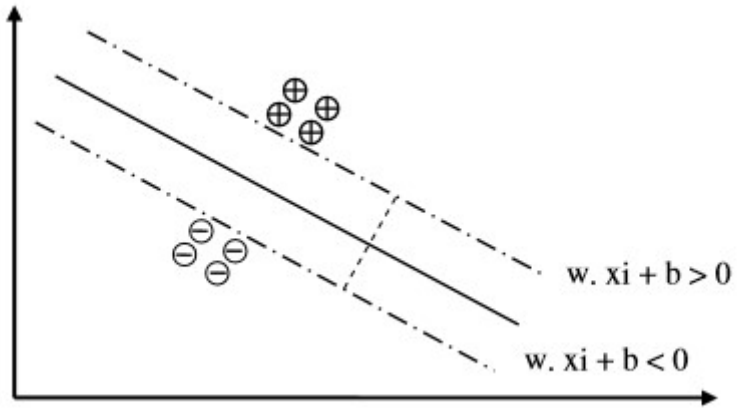


Figure 4.5: Support Vector Machine illustration

```
In [37]: from sklearn.svm import LinearSVC # import SVC model from sklearn.svm
         svm_clf = LinearSVC(random_state=0)

In [38]: svm_clf.fit(xtrain_bow, ytrain)

Out[38]: LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True,
                   intercept_scaling=1, loss='squared_hinge', max_iter=1000,
                   multi_class='ovr', penalty='l2', random_state=0, tol=0.0001,
                   verbose=0)
```

Confusion matrix, precision, recall, and F1 measures are the most usually utilized measurements for characterization errands. Scikit-Learn's measurements library contains the arrangement report and disarray lattice techniques, which can be promptly used to discover the qualities for these significant measurements.

```
In [42]: from sklearn.metrics import confusion_matrix # import confusion matrix from the sklearn.metrics
         confusion_matrix(yvalid, y_pred_svm)

Out[42]: array([[8808,   82],
                [ 409,  290]], dtype=int64)
```

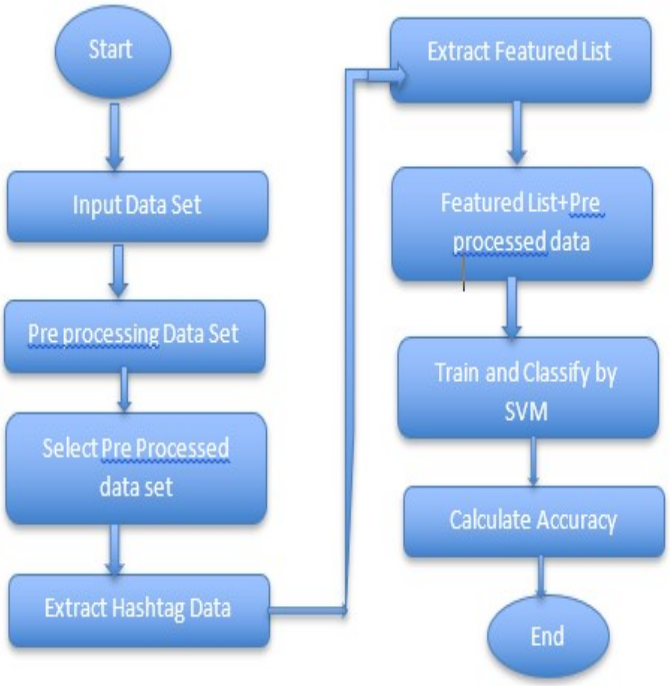


Figure 4.6: Full process of model evaluation

CHAPTER 5: RESULT ANALYSIS AND DISCUSSION

5.1 Recall, Precision, F-measure

I tried 9590 examples for estimation examination inquire about. There were 8808 True positive, 82 False positive, 409 False negative and 290 True negative information.

A perplexity framework is required for our anticipated outcome investigation. So now we will examine on precision, recall, and f-measure (Makhoul et al.,1999).

Disarray Lattice is an exhibition estimation for AI characterization. It is very valuable for estimating Recall, Accuracy and Precision.

Table 5. 1: Confusion matrix classes and parameters

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

True Positives (TP) ⇒ Translation: You anticipated positive and it's valid.

You anticipated that a man is visually impaired and he really is.

True Negatives (TN) ⇒ Elucidation: You anticipated negative and it's valid.

You anticipated that a man isn't hard of hearing and he really isn't.

False positives and b False negatives, those qualities happen when unique class denies with the anticipated class.

False Positives (FP) ⇒ Translation: You anticipated positive and it's bogus.

You anticipated that a lady is pregnant however she really isn't.

False Negatives (FN) ⇒ Elucidation: You anticipated negative and it's False.

You anticipated that a lady isn't deaf hearing yet she really is.

Table 5. 2: Naïve-Bayes metrics

True Positive	8915
False positive	2
True Negative	45
False negative	627

Table 5. 3: Logistic regression metrics

True Positive	8861
False positive	56
True Negative	248
False negative	424

Table 5. 5: SVM metrics

True Positive	8828
False positive	89
True Negative	281
False negative	391

Presently by these four parameters, we can compute Recall, Precission and F1 score for our expectation with SVM classifier.

Precision $\Rightarrow \frac{TP}{TP+FP}$

So the term is actually looks like -

$$Precision = \frac{8808}{8808+82}$$

$$= 0.990 \quad (5.1)$$

Recall $\Rightarrow \frac{TP}{TP+FN}$

So the term actually looks like-

$$Recall = \frac{8808}{8808+409}$$

$$= 0.956 \quad (5.2)$$

F-Measure : The F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

$$F\text{-}measure = \frac{2(recall*precision)}{recall+precision}$$

$$= \frac{2(0.990*0.956)}{0.990+0.956}$$

$$= 97 \quad (5.4)$$

So, accuracy of my proposed model is 96 percentage.

As I calculated precision, recall and f-measure with other algorithm, the result is-

```
In [46]: from sklearn.metrics import confusion_matrix
         confusion_matrix(yvalid, y_pred_lg)

Out[46]: array([[8861,   56],
                [ 424,  248]], dtype=int64)
```

```
In [49]: from sklearn.metrics import confusion_matrix
         confusion_matrix(yvalid, y_pred_nb)

Out[49]: array([[8915,    2],
                [ 627,   45]], dtype=int64)
```

```
In [41]: from sklearn.metrics import confusion_matrix
         confusion_matrix(yvalid, y_pred_svm)

Out[41]: array([[8828,   89],
                [ 391,  281]], dtype=int64)
```

Algorithm	Precision	Recall	F-measure
SVM	0.99	0.95	0.97
Naïve-Bayes	0.96	0.93	0.95
Logistic Regression	0.98	0.95	0.96

Support vector gave me highest value while Naïve-Bayes gave minimum value within these three.

CHAPTER 6: CONCLUSION

6.1 Our Contribution

Sentiment Analysis turns into the most significant source in basic leadership. Nearly individuals rely upon it to accomplish the proficient item. In spite of the fact that, there are countless people, who compose and readtwitter post day by day, the exploration in this field finds insufficient till now. Since breaking down logical posts space is hard. It has exceptional highlights and attributes impacts on the feeling extremity assessment. In this paper I Demonstrated the best model to characterize negative and positive information using hashtags on twitter information.

For assessing my proposed method I thought about four propelled AI calculations like SVM, Innocent Bayes, Greatest Entropy and Strategic Relapse. What's more, I have the most noteworthy exactness of SVM just as Strategic relapse. However, as a genuine order issue, fmeasure is better measurement to assess my model. And afterward Bolster vector machine gives the best f-measure than others. So I can say SVM got best outcome.

6.2 Future Works

In the near future, I intend to work on large dataset as support vector machine is not feasible for large dataset. While working on this paper I found the trick for training SVMs on large datasets is to approximate the optimization problem with a set of smaller subproblems. So next, (M, 2013)I intend to apply this trick on my work.

REFERENCES

[1] Louis, F.& Wojciech, C., Book "Advances in The Human Side of Service Engineering", 5th International Conference on Applied Human Factors and Ergonomics, Volume Set, Proceedings of the 5th AHEE Conference 19-23 July 2014.

[2] Larsen, Peder Olesen, and Markus von Ins. “The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index.”,Scientometrics 84.3 (2010): 575–603. PMC. Web. 25 Sept. 2015.

[3] Peng, L., Cui, G., Zhuang, M., and Li, C., “What do seller manipulations of online product reviews mean to consumers? ” (HKIBS Working Paper Series 070-1314). Hong Kong: Hong Kong Institute of Business Studies, Lingnan University, 2014.

[4] Thomas, B., Keep Social Honest , “What Consumers Think about brands on social media, and what businesses need to do about it” Report, 2013.

[5] Liu, B., Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[6] Yin, Z., Rong, J., and Zhi-Hua, Z., “Understanding Bag-of-Words Model: A Statistical Framework”, International Journal of Machine Learning and Cybernetics, 2010.

[7] Nisha, J., & Dr.E. Kirubakaran, “M-Learning sentiment analysis with Data Mining Techniques”, International Journal of Computer Science and Telecommunications, Volume 3, Issue 8, 2012.

[8] Richard, S., Alex, P., Jean, Y.W., Jason, C., Christopher, D.M., Andrew, Y.N. and Christopher, P., “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”, Conference on Empirical Methods in Natural Language Processing, 2013.

[9] Farah, B., Carmine, C., & Diego.R., “Sentiment analysis: Adjectives and Adverbs are better than Adjectives Alone”, International Conference on Weblogs and Social Media - ICWSM , Boulder, CO USA, 2007.

[10] Saifee, V., & Jay, T., “Applications and Challenges for Sentiment Analysis: A Survey”, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue2, 2013.

[11] Hassena, R.P., “Challenges and Applications”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 5, 2014.

[12] Apoorv, A., Boyi, X., Ilia, V., Owen, R., and Rebecca, P.,"sentiment Analysis of Twitter Data", proceeding LSM '11 Proceedings of the Workshop on Languages in Social Media, Pages 30-38, USA, 2011. [13] Anuj, S. & Shubhamoy, D.,"Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis",Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012 [14] Andranik, T., Timm, O., Philip, G.S., and Isabel, M.W., "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, Association for the advacement of artifical intelligence 2010 [15] Mitchell, P.M., Mary A.M., and Beatrice, S., "Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics Journal , Vol. 19, Number 21993.

[16] Schukla, A., "Sentiment analysis of document based on annotation", CORR Journal, Vol. abs/1111.1648, 2011.

[17] Kasper, W. & Vela, M., "Sentiment analysis for hotel reviews", proceedings of the computational linguistics-applications, Jacharanka Conference, 2011.

[18] Zhang, L., Hua, K., Wang, H., and Qian, G.,"Sentiments reviews for mobile devices products", The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2014) ,procedia computer scinence, Volume 34, 2014.

[19] Godbole, N., Srinivasaiah, M., and Skiena, S., "Large-Scale Sentiment Analysis for News and Blogs", ICWSM'2007 Boulder, Colorado, USA, 2007.

[20] Esuli A., & Sebastiani, F.,"SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining", Kluwer Academic Publishers. Printed in the Netherlands, 2006.

[21] Hearst, M., "Direction-based text interpretation as an information access refinement", In Paul Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates, 1992.

[22] Das, S., and Chen, M., “Yahoo! for Amazon: Extracting market sentiment from stock message boards”, In Proc. of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001), 2001.

[23] Turney, P., “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”. In Proc. of the ACL, 2002.

[24] Argamon-Engelson, S., Koppel, M., and Avneri, G., “Style-based text categorization: What newspaper am I reading? ”, In Proc. of the AAAI Workshop on Text Categorization, pages 1–4, 1998.

[25] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D. and Yates, A., “Unsupervised named-entity extraction from the web: An experimental study”, Artificial Intelligence, 165(1):91–134, 2005.

[26] Pang, B. & Lee, L., “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”, Association of Computational Linguistics (ACL), 2004.

[27] Jin, W., & HO H. H., “A novel lexicalized HMM-based learning framework for web opinion mining”, Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada, ACM: 465-472, 2009.

[28] Brody, S., & Elhadad, N., “An unsupervised aspect-sentiment model for online reviews”, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, California, Association for Computational Linguistics: 804-812, 2010.

[29] Wiebe, J., Wilson, T., and Cardie, C., “Annotating expressions of opinions and emotions in language”. Language Resources and Evaluation, 39(2-3):165–210, 2005.

[30] Cui, A., Zhang, M., Liu, Y., and Ma, S., “Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis”, In Information Retrieval Technology, pages 238–249. Springer, 2011.

[31] Kumar, A., & Sebastian, T. M., “Sentiment Analysis On Twitter”. IJCSI International Journal of Computer Science, Issues 9(3):372–378, 2012.

[32] Nielsen, F. A. “ANEW: Evaluation of a word list for sentiment analysis in microblogs”. In Proceedings of the ESWC2011 Workshop on ‘Making Sense of Microposts’: Big things come in small packages, volume 718 of CEUR Workshop Proceedings, pages 93–98, 2011.

[33] Minocha, A. & Singh, N., “Generating domain specific sentiment lexicons using the web directory”. Advanced Computing: an International Journal, 3, 2012.

[34] Kamps, M., Marx, R.J. Mokken, and Rijke, M.D., “Using wordnet to measure semantic orientation of adjectives”, procedding in National Institute for (2004) 1115– 1118, 2004.

[35] Kim, S., & Hovy, E., “Identifying and analyzing judgment opinions”. In: Proceedings of HLT/NAACL-2006. (2006) 200–207.

[36] Rao, D., & Ravichandran, D., “Semi-supervised polarity lexicon induction”. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. EACL ’09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 675–682.

[37] Carmen Banea, R.M., & Wiebe, J., “A bootstrapping method for building subjectivity lexicons for languages with scarce resources”, In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), Marrakech, Morocco, European Language Resources Association (ELRA) , May 2008.

[38] Radulescu, C., Dinsoreanu, M., and Potolea, R., “Identification of spam comments using natural language processing techniques”, Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on, Sepbtemebr 2014.

[39] Zhang, Q., Wu, Y., Li, T., Ogihara, M., Johnson, J., and Huang, X., “Mining Product Reviews Based on Shallow Dependency Parsing”, SIGIR’09, July 19–23, Boston, Massachusetts, USA. ACM 978-1-60558-483-6/09/07, 2009.

[40] Emus, R., “Modeling and Representing Negation in Data-driven Machine Learning-based Sentiment Analysis”, In Proceedings of the 1st International

[41] R. Feldman, ” Techniques and Applications for Sentiment Analysis ,” Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.

[42] Y. Singh, P. K. Bhatia, and O.P. Sangwan, ”A Review of Studies on Machine Learning Techniques,” International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 70-84, 2007.

[43] P.D. Turney,” Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, July 2002.

[44] Ch.L.Liu, W.H. Hsaio, C.H. Lee,and G.C.Lu, and E. Jou,” Movie Rating and Review Summarization in Mobile Environment,” IEEE Transactions on Systems, Man, and Cybernetics, Part C 42(3):pp.397-407, 2012.

[45] Y.Luo,W.Huang,” Product Review Information Extraction Based on Adjective Opinion Words,” Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1309 – 1313, 2011.

[46] R.Liu,R.Xiong,and L.Song, ”A Sentiment Classification Method for Chinese Document,” Processed of the 5th International Conference on Computer Science and Education (ICCSE), pp. 918 – 922, 2010.

[47] A.khan,B.Baharudin, ”Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs,” Processed on National Postgraduate Conference (NPC), pp. 1 – 7, 2011.

[48] L.Ramachandran,E.F.Gehringer, ”Automated Assessment of Review Quality Using Latent Semantic Analysis,” ICALT, IEEE Computer Society, pp. 136-138, 2011.

[49] B.Agarwal,V.K.Sharma,andN.Mittal,”Sentiment Classification of Review Documents using Phrase Patterns,” International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1577-1580, . 2013.

[50] J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,,” Aspect-Based Opinion Polling from Customer Reviews,” T. Affective Computing2(1):pp. 3749, 2011.

[51] M.Karamibekr,A.A.Ghorbani,”Verb Oriented Sentiment Classification,” Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp. 327-331, 2012.

[52] A. Neviarouskaya, H.Prendinger, and M.Ishizuka,” SentiFul: A Lexicon for Sentiment Analysis,” T. Affective Computing 2(1), pp.22-36, 2011.

[53] L.Liu, X.Nie,and H.Wang,” Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis,” Processed of the 5th Image International Congress on Signal Processing (CISP), pp. 1620 – 1624, 2012.

[54] R. Srivastava, M. P. S. Bhatia," Quantifying Modified Opinion Strength: A Fuzzy Inference System for Sentiment Analysis," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1512-1519, 2013.

[55] C. Tillmann , and F. Xia, "A phrase-based unigram model for statistical machine translation," Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL, pp.106-108, 2003.

[56] B.Ren,L.Cheng," Research of Classification System based on Naive Bayes and MetaClass," Second International Conference on Information and Computing Science, ICIC '09, Vol(3), pp. 154 – 156, 2009.

[57] C.I.Tsatsoulis,M.Hofmann,"Focusing on Maximum Entropy Classification of Lyrics by Tom Waits," IEEE International on Advance Computing Conference (IACC), pp. 664 – 667, 2014.

[58] M.A. Hearst,"Support vector machines,"IEEE Intelligent Systems, pp. 18-28, 1998.

[59] R. Feldman, "Techniques and applications for sentiment analysis," Commun. ACM, vol. 56, no. 4, pp. 82–89, Apr. 2013.

[60] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Computer Science, vol. 17, no. 0, pp. 26 – 32, 2013, first International Conference on Information Technology and Quantitative Management.

[61] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," Expert Systems with Applications, vol. 36, no. 7, pp. 10760 – 10773, 2009.

[62] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," J. Am. Soc. Inf. Sci. Technol., vol. 62, no. 2, pp. 406–418, Feb. 2011.

[63] M. M. Mostafa, “More than words: Social networks’ text mining for consumer brand sentiments,” Expert Systems with Applications, vol. 40, no. 10, pp. 4241 – 4251, 2013.

[64] L. C. Borges, V. M. Marques, and J. Bernardino, “Comparison of data mining techniques and tools for data classification,” in Proceedings of the International C* Conference on Computer Science and Software Engineering, ser. C3S2E ’13. New York, NY, USA: ACM, 2013, pp. 113–116.

[65] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proceedings of the ACL, 2004.

[66] M. Taboada, C. Anthony, and K. Voll, “Methods for creating semantic orientation dictionaries,” in Conference on Language Resources and Evaluation (LREC), 2006, pp. 427–432.

[67] 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, ser. LSM ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 30–38.

[68] T. S. Zakzouk and H. I. Mathkour, “Comparing text classifiers for sports news,” Procedia Technology, vol. 1, no. 0, pp. 474 – 480, 2012, first World Conference on Innovation and Computer Sciences (INSODE 2011).

[79] M. R. Saleh, M. Mart´ın-Valdivia, A. Montejo-R´aez, and L. Ure˜naL´opez, “Experiments with svm to classify opinions in different domains,” Expert Systems with Applications, vol. 38, no. 12, pp. 14799 – 14804, 2011.

[80] M. Abdul-Mageed, M. Diab, and S. K¨ubler, “Samar: Subjectivity and sentiment analysis for arabic social media,” Computer Speech & Language, vol. 28, no. 1, pp. 20 – 37, 2014.

[81] A. Abbasi, H. Chen, and A. Salem, “Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums,” ACM Trans. Inf. Syst., vol. 26, no. 3, pp. 12:1–12:34, Jun. 2008.

[82] A. K. Uysal and S. Gunal, “A novel probabilistic feature selection method for text classification,” Know.-Based Syst., vol. 36, pp. 226–235, Dec. 2012.

[83] M. F. Porter, “Readings in information retrieval,” K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. An Algorithm for Suffix Stripping, pp. 313–316.

[84] S. Gunal and R. Edizkan, “Subspace based feature selection for pattern recognition,” Information Sciences, vol. 178, no. 19, pp. 3716–3726, 2008.

[85] T. L.Ladha, “Feature selection methods and algorithms,” International Journal on Computer Science and Engineering (IJCSE), vol. 3, p. 5, 2011.

[86] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” Information Processing & Management, vol. 50, no. 1, pp. 104 – 112, 2014.

[87] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[89] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in Machine Learning: ECML-98, ser. Lecture Notes in Computer Science, C. N´edellec and C. Rouveirol, Eds. Springer Berlin Heidelberg, 1998, vol. 1398, pp. 137–142.

[90] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: Sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 Conference on Empirical

Methods in Natural Language Processing - Volume 10, ser. EMNLP ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[91] E. Martinez Camara, V. Martin, M. Teresa, J. M. Perea Ortega, and L. A. Urena Lopez, “Tecnicas de clasificacion de opiniones aplicadas a un corpus en espanol,” Procesamiento de Lenguaje Natural, vol. 47, pp. 163–170, 2011.

[92] F. Sebastiani, “Machine learning in automated text categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, Mar. 2002.

[93] Doaa Mohey El-Din, "Negative Sentiment Analysis Polarity Levels", at Future Technologies Conference 2016, The Science and Information (SAI) Organization Thomson Reuters approved organization, published IEEE Xplore.

