## North Western University, Khulna

**Computer Science and Engineering**

The undersigned hereby certify that they have read and recommend to the Computer Science and Engineering for acceptance a thesis entitled “**HUMAN EMOTION ANALYSIS VIA TWITTER**” by SK. Tanjis Hossain, Md. Habibur Rahman, Anindya Sundor Roy and Sumiya Jahan Linain partial fulfillment of the requirements for the degree of **Bachelor of Science in Computer Science and Engineering (CSE).**

Dated: 22th December, 2018

Md. Mahedi Hasan Thesis Supervisor

Lecturer

Computer Science and Engineering Discipline

North Western University, Khulna.

Nasif Alvi Second Examiner

Lecturer

Computer Science and Engineering Discipline

North Western University, Khulna.

Nagib Mahfuz Head of the

Lecturer

Computer Science and Engineering Discipline

North Western University, Khulna.

# Acknowledgements

First of all, we are grateful to Almighty ALLAH the merciful, the gracious who has given us the ability, intelligence and energy to complete our thesis.

We would like to express our most sincere gratitude to our supervisor Lecturer Md. Mahedi Hasan for his extraordinary patience, continuous encouragement, guidance and advice. We have been amazingly fortunate to have an advisor who gave us the freedom to explore on our own. This research experiment would not have been possible without his support.

We would like to thank our Second Examiner Lecturer Nasif Alvi for his proper guidelines and reviews of our work. We are deeply grateful to him for his support and the long discussions that helped us to sort out the technical details of our work.

Moreover, we would like to thank all our respected teachers of CSE discipline for their insightful comments, encouragement and valuable advice. We thank our classmates for stimulating discussions, for the sleepless nights we were working together before deadlines and for all the fun we have had in the last four years. We thank our seniors for their wise, valuable information and heartiest cooperation. We are also thankful to our juniors for their goodwill support. Whenever we met, they used to ask about our thesis progress.

Last but not the least, we would like to thank our family, our parents and to our brothers for supporting us spiritually throughout writing this thesis and our life in general.

# Abstract

Twitter is an online social networking service on which users worldwide publish their opinions on a variety of topics, discuss current issues, complain, and express many kinds of emotions. Therefore, Twitter is a rich source of data for opinion mining, sentiment and emotion analysis. Sometimes it is difficult to understand the user’s opinion. The main challenge is to feature extraction for the purpose of classification and feature extraction depends on the perfection of preprocessing of a tweet. The preprocessing is the most difficult task, since it can be done in various ways and the methods or steps applied in preprocessing are not distinct. Most of the researches in this topic, has focused on binary (positive and negative), 3-way (positive, negative and neutral) and 4-way (happiness, sadness, surprise, disgust) classifications. In this paper, we focus on emotion classification of tweets as multi-class classification. We have chosen basic human emotions (happiness, sadness, love, hate, surprise and worry) classification. According to the experimental results, our approach improved the performance of multi-class classification of twitter data.

# Contents

[Chapter 1](#_bookmark0)

[Introduction 1](#_bookmark1)

* 1. [Motivation](#_bookmark3) 2
  2. [Sentiment Analysis](#_bookmark4) 2
  3. [Sentiment Analysis vs. Emotion Analysis](#_bookmark5) 2
  4. [Objectives](#_bookmark6) 3
  5. [Thesis Organization](#_bookmark7) 3

[Chapter 2](#_bookmark8)

[Study Survey](#_bookmark9) 4

* 1. [Raw Data](#_bookmark12) 4
  2. [Preprocessing](#_bookmark14) 4
  3. [Feature Extraction](#_bookmark15) 4
  4. [Related Work](#_bookmark17) 5

[Chapter 3](#_bookmark18)

[Methodology](#_bookmark19) 9

* 1. [Data Preparation](#_bookmark21) 9

3.2 Preprocessing 9

3.2.1 Removing Emoticons and Symbol 9

3.2.2 Word Correction 9

3.2.3 Punctuation Removal 9

3.2.4 @username Removal 10

3.2.5 Hashtag Removal 10

3.2.6 Stop Word Removal 10

3.2.7 URL (Uniform Resource Locator) Removal 10

3.2.8 Stemming and Lemmatization 10

3.3 [Methodology 1](#_bookmark22)0

3.3.1 Waikato Environment for Knowledge Analysis (Weka) 10

3.3.2 Weka Data Formats 11

3.3.3 Weka Explorer 11

3..3.4 Cross Validation Fold 12

* + 1. Split Percentage 12

3.3.6 Attribute-Relation File Format (ARFF) 12

* 1. [Feature Extraction Model 1](#_bookmark23)2
     1. [Unigram Model 1](#_bookmark24)3
     2. N[aïve Bayes 1](#_bookmark36)3

3.4.3 Multinomial Naïve Bayes 14

[Chapter 4](#_bookmark43)

[Evaluation and Results 1](#_bookmark44)5

* 1. [Data Set 1](#_bookmark45)5
  2. [Tools Used for Our Thesis](#_bookmark49) 15
     1. [Microsoft Excel and Text Document](#_bookmark50) 15
     2. [Preparing Data For Classification](#_bookmark51) 15
  3. [Experimental Results](#_bookmark55) 16
  4. Analysis 24

[Chapter 5](#_bookmark69)

[Conclusions and Future Directions 2](#_bookmark70)5

* 1. [Conclusion 2](#_bookmark71)5
  2. [Limitations 2](#_bookmark72)5
  3. [Future Directions 2](#_bookmark73)5

**List of Figure**

**Figure: Name of Figure Page**

[Figure 1.1: Number of monthly active twitter users in millions](#_bookmark2) 1

[Figure 4.1: ARRF File 1](#_bookmark20)6

[Figure 4.2: WEKA explorer preprocess tab](#_bookmark20) 17

[Figure 4.3: Edit tab viewer](#_bookmark25) 17

[Figure 4.4: classifier evaluation options](#_bookmark58) 18

[Figure 4.5: weka explorer classify layout](#_bookmark63) 19

[Figure 4.6: weka building time](#_bookmark20) 19

[Figure 4.7: weka accuracy rate](#_bookmark25) 19

[Figure 4.8: Accuracy of emotion classes for 6-way classification](#_bookmark58) 20

[Figure 4.9: Accuracy of emotion classes for 2-way classification](#_bookmark63) 21

[Figure 4.10: weka confusion matrix](#_bookmark20) 22

[Figure 4.11: weka detailed Accuracy by class](#_bookmark25) 22

[Figure 4.12: weka visualization of given dataset](#_bookmark58) 23

[Figure 4.13: weka independent frequency of a class](#_bookmark58) 23

## List of Tables

**Table: Name of Table Page**

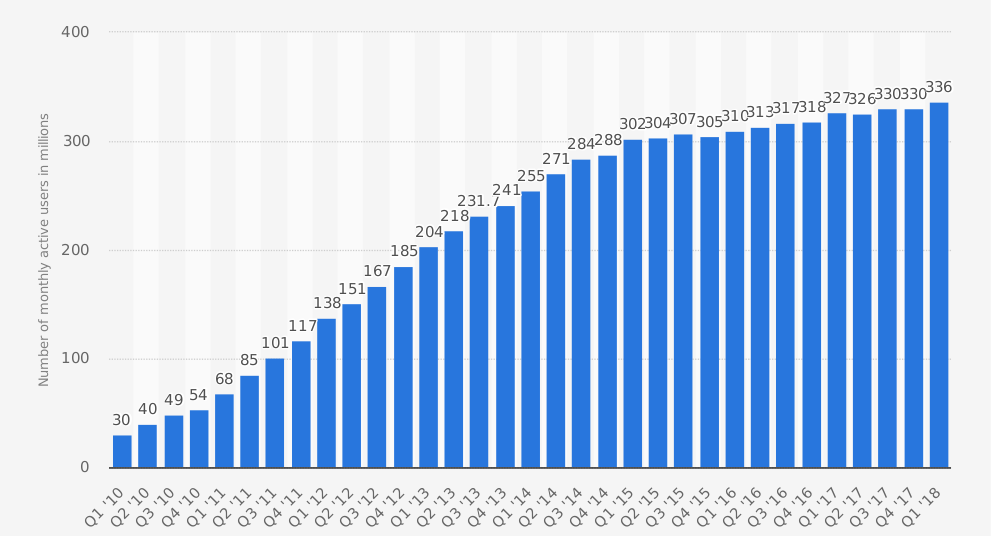
[Table 1: Test accuracy and average accuracy for 6-way classification](#_bookmark26) 20

[Table 2: Test accuracy and average accuracy for 2-way classification](#_bookmark27) 21

[Table 3: Comparison of accuracy of our system and Abu et al 2](#_bookmark57)4

**Chapter 1**

# **Introduction**

Microblogging websites such as Twitter (www.twitter.com) have evolved to become a great source of various kinds of information. This is due to the nature of microblogs on which people post real-time messages regarding their opinions on a variety of topics, discuss current issues, complain, and express many kinds of emotions. As the audience of microblogging platforms and social networks grows every day, data from these sources can be used in opinion mining, sentiment and emotion analysis tasks. Opinions and related concepts such as sentiments and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, microblogs, Twitter, and social networks. Almost all forms of social media are very noisy and full of all kinds of spelling, grammatical, and punctuation errors. In our proposition, we have concentrated on twitter, the most well-known microblogging stage. The quantity of twitter clients achieved an expected 336 million in this year, up from around 6 million in earlier year. Figure 1 speaks to the quantity of month to month dynamic clients of twitter in millions from first quarter of 2010 to initially quarter of 2018 where each quarter speaks to 3 months of a year. Tweets are more casual and limited to 140 characters of text but twitter has announced plans to double the character limit allowed in an individual tweet from the familiar 140 characters up to 280[22]. Twitter contains countless messages. Regular clients as well as superstars, organization delegates, lawmakers and even nation presidents are crowds of twitter.

**Figure 1.1: Number of monthly active twitter users in millions [12]**

In our thesis, we have applied a technique for emotion analysis of tweets using unigram model for feature extraction and Naïve Bayes classifier for classification of emotions. We have classified tweets into six classes (happiness, sadness, surprise, worry, hate and love) where six classes of emotion have been chosen according to the theories of psychologist Paul Ekman’s basic human emotions for our research. Most of the previous researches in this area used binary classification (positive, negative),3-way classification (positive, negative and neutral) and 4-way classification(happy, sad, surprise, disgust). Here we have tried to achieve a new aspect of multiclass classification of human emotion.

## Motivation

Among internet users Microblogging such as Twitter has become a very popular communication tool where users share opinion in different aspects of life. People share their thoughts about products and services. That’s why we can consider Twitter as a valuable source of information for opinion mining and sentiment analysis. Due to the rapid grown of the social media, the amount of customer feedback has increased and is available to corporations, business owners and different kind of companies. These companies are interested obtaining customer feedback. So we have chosen to work with twitter since it provides a better approximation of public opinion or sentiment than internet articles, reviews or web blogs. Moreover, the response in twitter is more rapid and also more general since the number of users in twitter is significantly more than the number of web bloggers on a daily basis. Analyzing twitter data, we can find the topics of interest or point of view on discussion. Customer service can be improved by finding dissatisfaction or problem with products. Using sentiment analysis result we can optimize survey.

## Sentiment Analysis

Sentiment analysis which is also known as opinion mining is an approach to identify subjective information in source material. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics. The goal of sentiment analysis is to find the attitude of a person with respect to some specific topic.

## 1.3 Sentiment Analysis vs. Emotion Analysis

Sentiment can be expressed mostly in binary format (positive and negative) or sometimes it can be expressed as tri-polar (positive, negative and neutral). However, emotions have multi- dimensions while the common generic emotions are happiness, sadness, anger, fear, disgust, surprise. Anger is for sure negative sentiment and so is sadness, but aren’t really same. We can say, emotion analysis can be referred to as next level of sentiment analysis.

* 1. **Objectives**

The major objectives of our thesis

* + - To collect tweets for training data and manually annotate the tweets into separate classes.
    - To process the data to achieve better result in classification.
    - To determine whether a tweet represents happiness, sadness, surprise, worry, hate and love.
    - To use different models to determine emotion of the tweets and to compare the accuracy of different models.

**1.5 Thesis Organization**

This thesis has been organized into five chapters. Each chapter gives distinct concept.

Chapter 1 (Introduction): Introduction of our research area has been explored in this section.

Chapter 2 (Study Survey): This chapter presents the related works about our thesis.

Chapter 3 (Methodology): The chapter presents the main architecture of the proposed system.

Chapter 4 (Evaluation and Results): Our experimental procedure and results are represented in this section.

Chapter 5 (Conclusions and Future Directions): Summarization of our research work. Some limitations and future plan of our research is also included.

**Chapter 2**

## Study Survey

In recent years, huge measure of inquires about have been done on feeling examination of twitter information. Most examinations started with gathering the ideal datasets shape twitter, and connected different sifting methods to evacuate excess information. At that point parsed the information into an organized shape to locate the correct highlights and broke down the information. Beneath we survey a few kinds of investigation that most explores have utilized on twitter information.

## 2.1 Raw Data

An essential part of sentiment analysis is to have a comprehensive dataset to train a model. Presently a-days there remains a huge resource of tweets or twitter data in online. There are many ways of getting twitter APIs to get tweets. Twitter itself allows user their own archive of posts and save them in an exported format [13] and there are also several tools [14] to get tweets such as Hootsuite, BirdSong Analytics, Cyfe, NodeXL, TWchat, TweetStats, Twittonomy, Tweettreach. After that these tweets can be manually or heuristically or automatically annotated to use as training dataset.

## 2.2 Preprocessing

Now a days Tweets are restricted to 280-characters length with an exceptionally easygoing language and cases tweets may be very noisy with username, links, repeated letters and emoticons. So different types of preprocessing on the preparation informational collection have been be finished. Like expelling non-English tweets, URL, target makes reference to, hashtags, numbers, relational words, stop-words and so forth. Specialist can supplant emojis by their extremity, negative notices by some invalidation tag, succession of rehashed characters by an explicit number of characters, etc. Tokenization, Normalization, Stemming and lemmatization are likewise utilized in the means of preprocessing. These preprocessing thoroughly rely upon the specialists dependent on their motivation of the examination.

## 2.3 Feature Extraction

This is the best part in microblogging sentiment analysis. Since people express opinions in complex ways. Most of the tweets are unstructured and also non-grammatical. And there remains a lot of lexical variation and extensive usages of acronyms like bcz,tc, gdn8, gdm9,asap, lol, btw etc. Researchers have used variety of features for their classification experiments on sentiment analysis.

For example:

Sample sequence: “John loves to play cricket games”

Unigram: “John”, “loves”, “to”, “play”, “football”, “games”

Bigram: “John likes”, “likes to”, “to watch”, “watch football”, “football games”

Trigram: “John likes to”, “likes to watch”, “to watch football”, “watch football games”

## 2.4 Related Works

Different researchers have proposed distinctive techniques for slant investigation. Some dependent on vocabulary highlights or probabilistic, some dependent on machine learning systems, and some dependent on joined those two procedures. They are quickly talked about beneath.

Sentiment analysis, or opinion mining, is the computational study of people’s opinions, sentiments, emotions, and attitudes. It is one of the most active research areas in natural language processing and is also extensively studied in data mining, web mining, and text mining. The growing importance of sentiment analysis coincides with the growth of social media, such as Twitter, Facebook, book reviews, forum discussions, blogs, etc. The basis of many sentiment-analysis approaches is the sentiment lexicons, with the words and phrases classified as conveying positive or negative sentiments. Several general-purpose lexicons of subjectivity and sentiment have been constructed.

In Abu et al [2] built model for four and five classification tasks: a 4-way task of classifying sentiment into happy, sad, surprise and disgust classes and a 5-way task of classifying sentiment into happy, sad, surprise, disgust and neutral classes. They experimented with two types of models: unigram model and unigram with POS tagging model. They used naïve bayes classifier to classify the data. They achieved highest result in 4 way classification amounting 81% on unigram and 79.5% unigram POS tagging individually, but in 5 way classification they got 66% on unigram and 64.8% on unigram POS tagging separately.so in this case they decided to eliminate 5 way classification because of their poor result and on the other hand their neutral class result was 20% and 18% . In their survey, the authors describes the existence techniques and approaches for a human opinion-oriented information retrieval. For future work they believed that the accuracy could still be improved and they desired to work with six classification.

Agarawal et al [1] built model for two classification tasks: a binary task of classifying sentiment into positive and negative classes and a 3-way task of classifying sentiment into positive, negative and neutral classes. They experimented with three types of models: unigram model, a feature based model and a tree kernel based model. They used manually annotated twitter data where each tweet is labeled as positive, negative, neutral or junk and they eliminated tweet with junk label for experiments

Pak and Paroubek [3] collected a corpus of 300000 tweets from Twitter and evenly splitted these into three sets of texts: texts containing positive emotions, negative emotions and no emotions. They queried two types of emoticons: happy emoticons and sad emoticons. They assumed an emoticon within a message represents an emotion for whole message and all the word of the message are related to this emotion. For corpus analysis they checked distribution of words frequencies in the corpus. They used TreeTagger for English to tag all the posts in the corpus.

In [6] (Yang et al., 2007), the authors use web-blogs to construct a corpora for sentiment analysis and use emotion icons assigned to blog posts as indicators of users’ mood. The authors applied SVM and CRF learners to classify sentiments at the sentence level and then investigated several strategies to determine the overall sentiment of the document. As the result, the winning strategy is defined by considering the sentiment of the last sentence of the document as the sentiment at the document level.

Balabantaray et al [16] attempted to characterize six fundamental feelings. They gathered tweets from web. The asset was set up by downloading 1000 arbitrarily chosen twitter client and their tweets. For information explanation they utilized five judges. Each sentence is exposed to two judgments. For learning model they utilized SMV bit. They utilized many element extraction models like unigram, bigram, POS, Word-net Affect feeling vocabulary and so on. They got in general 73.24% exactness result.

Purver and Battersby [15] in their exploration they for the most part took a shot at emojis and hashtags. In their work, they ordered emojis and hashtags with six essential feelings. Characterization in all analyses they utilized help vector machines (SVM) by means of LIBSVM execution with direct part and unigram highlights. In this paper three individual tests were finished. Trial 1: Emotion identification, test 2: feeling separation and test 3: manual marking. These trials depend on the two emojis and hashtags and they got 50-80% precision shifts from investigation to try. Further they set up a web to look at whether dependably arrange these emojis. The review is finished by 492 people.

Go et al [24] in their research didn’t consider neutral tweets in their training or testing data. They only used positive and negative tweets. Their approach was to use machine learning classifier and feature extractors. They stripped the emoticons out from their training data because according to their research emoticons does negative impact on the accuracy of Maximum Entropy and SVM classifier, but little impact on Naïve Bayes. In feature reduction process, an equivalent class token(“USERNAME”) was replaced all words that start with the “@” symbol. They converted URL like <http://tinyurl.com/cvvg9a> to the token “URL”. They used preprocessing so that any letter occurring more than two times in a row is replaced with two occurrences. They tested different classifiers: key-based, Naïve Bayes, Maximum Entropy, and support vector machines. For the training data, they used scrapper that queries the Twitter API. In their paper, the training data was post-processed with some filters like stripping off the emoticons and removing of retweets. They explored the usage of unigrams, bigrams, unigrams and bigrams, and parts of speech as features. They showed that unigrams and bigrams gave the best result. For future work they believed that the accuracy could still be improved.

Barbosa and Feng [5] proposed a 2-step sentiment analysis classification method for twitter, which first classifies message as subjective and objective and further distinguished the subjective tweets as positive and negative. They proposed the use of two sets of features: meta-information about the words on tweets and characteristics of how tweets are written. They created a single classifier combined objectivity sentence from Twendz at <http://twendz.waggeneredstrom.com/> and Twitter Sentiment at <http://twittersentiment.appspot.com/> (objectivity class) and subjectivity sentences from all the three sources (TweetFeel at <http://www.tweetfeel.com/> ). They tried different learning algorithms available on Weka ( Witten and Frank,2005) and SVM obtained best result for unigram and TwitterSA.

Most sentiment-analysis [7] research focuses on English text and, consequently, most of the resources developed are in English. Emoticons have proved crucial in the automated sentiment classification of informal texts. In an early work, a basic distinction between positive and negative emoticons was used to automatically generate positive and negative samples of texts. These samples were then used to train and test sentiment-classification models using machine learning techniques.

Emoji’s, a new generation of emoticons, are increasingly being used in social media. Tweets, blogs and comments are analyzed to estimate the emotional attitude of a large fraction of the population to various issues. An emoji sentiment lexicon, provided as a result of this study, is a valuable resource for automated sentiment analysis. The Emoji Sentiment Ranking has a format similar to SentiWordNet, a publicly available resource for opinion mining, used in more than 700 applications and studies so far, according to Google Scholar. In addition to a public resource, the paper provides an in-depth analysis of several aspects of emoji sentiment.

Accordingly, we ask what are the different activities on Twitter that represent influence of a user and to what extent a person’s influence varies across tweet topic and time. We describe how we collected the Twitter data and present the characteristics of the top users based on three influence measures: in degree, retweets, and mentions. The social link information is based on the final snapshot of the network topology at the time of crawling and we do not know when the links were formed.

The network of Twitter users comprises a single disproportionately large connected component (containing 94.8% of users), singletons (5%), and smaller components (0.2%). The largest component contains 99% of all links and tweets. Our goal is to explore influence of users, hence we focus on the largest component of the network, which is conceptually a single interaction domain for users.

Since it is difficult to decide impact of clients who have few tweets, we obtained the idea of "dynamic clients" from the conventional media look into sand concentrated on those clients with some base level of action. We overlooked clients who had posted less than 10 tweets amid their whole lifetime.

We additionally disregarded clients for whom we didn't have a substantial screen name, since this data is critical in distinguishing the occasions a client was made reference to or retweeted by others. In the wake of sifting, there were 6,189,636 clients, whom we center around in the rest of this paper. To quantify the impact of these 6 million clients, be that as it may, we investigated how the whole arrangement of 52 million clients communicated with these dynamic clients.

**Chapter 3**

# **Methodology**

## 3.1 Data Preparation

As sentiment analysis of tweets is a very popular topic in research nowadays, there remains a lot of labeled data in online. Twitter APIs can also be used to access twitter data. Though there remain several papers on emotion analysis, but no labeled dataset of emotion analysis cannot be found in online. So we have to label the data by ourselves. The website figure-eight provides us with a human labeled corpus with over 40001 tweets with thirteen polarities (anger, boredom, empty, enthusiasm, fun, happiness, hate, love, neutral, relief, sadness, surprise, worry). We have collected our tweets from this site then labeled manually.

## 3.2 Preprocessing

After getting those labeled data we have processed the data because tweets contain a lot of noisy data because of its short length. We have removed the emoticons, URLs, targets, punctuation, stop- words. We have converted the whole tweet or text into small letters and also have applied stemming to identify a word by its root. Here we discuss about the correction of data. Figure 4 represents the steps of our preprocessing.

## Removing Emoticons and Symbols

In the second step of preprocessing we have removed the emoticons and different kind of symbols used in the tweets. We have seen that some researchers used emoticons in their experiment. They used the polarity of the emoticons to detect the sentiment or opinion of the tweet. But we want to identify the emotion of a tweet by using only the text. That’s why we have removed the emoticons. As tweets are very noisy, there remains various kind of symbols which are used by the users. Those symbols represent nothing, that’s why we have removed these different kind of symbols. For example, “no new movie for 3 months. \*sniffs\* I'm going to be so bored .”, here the symbol “\*”, “\*” will be removed and the emoticon “” will also be deleted.

## Word Correction

In tweets there are sequence of repeated characters like “cooooooooool”, “wishhhhhhhh”, “happyyyyyyyyyyy”, “yesssssssssssss”, etc. To correct the words like these we have converted the repeated sequence of characters into two characters like “cool”, “wishh”, “happyy” ,“yes”.

## Punctuation Removal

In this section of preprocessing we have removed the punctuation marks. Because text with punctuation doesn’t represent any sentiment or emotions.

## @username Removal

The “@” sign is used for mention usernames in tweets [21]. Usually people use @username to mention a user in tweets. We removed these user mention because it doesn’t mean any kind of emotions.

## Hashtag Removal

By clicking on the hashtag users can view other tweets containing the same keyword or topic [21]. We have seen that some researches were done using hashtag. But we have mentioned that we want to work on only human emotions based on text, we have removed all the hashtags from the tweets.

## Stop Word Removal

Stop word are usually referred to as the most common used words in a language. Such stop words are “to”,”it”, “if”, “their”, “that”, “with”, “into”, etc. There is no specific list or limitation of stop words. We removed these stop words from our tweets for getting a better result from our research.

## URL (Uniform Resource Locator) Removal

In out collected tweets we find many URL like @username, other website links, media links, etc. This URL doesn’t mean any kind of sentiments. It just links to other user profiles, websites. That’s why we removed all kind of links or URLs from our tweets.

## Stemming and Lemmatization

Stemming and Lemmatization [19] are Text Normalization or sometimes called Word Normalization techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing. The goal of both stemming and lemmatization [20] is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am, are, is $\Rightarrow$ be

cars, car's, cars' $\Rightarrow$ car

By the following rules the result is:

the boy's cars are different colors $\Rightarrow$ the boy car be differ color

## 3.3 Methodology

## 3.3.1 Waikato Environment for Knowledge Analysis (Weka)

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, NewZealand. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

* Free availability under the GNU General Public License.
* Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
* A comprehensive collection of data preprocessing and modeling techniques.
* Ease of use due to its graphical user interfaces.

## 3.3.2 Weka Data Formats

## Weka [8] uses the Attribute Relation File Format for data analysis, by default. But listed below are some formats that Weka supports [23], from where data can be imported:

* CSV
* ARFF
* Database using ODBC

## 3.3.3 Weka Explorer

The Weka Explorer [9] is as follows:  
i) Preprocess: This allows us to choose the data file.  
ii) Classify: This allows us to apply and experiment with different algorithms on preprocessed data files.  
iii) Cluster: This allows us to apply different clustering tools, which identify clusters within the data file.  
iv) Association: This allows us to apply association rules, which identify the association within the data.  
v) Select attributes:These allow us to see the changes on the inclusion and exclusion of attributes from the experiment.  
vi) Visualize: This allows us to see the possible visualization produced on the data set in a 2D format, in scatter plot and bar graph output.

The user cannot move between the different tabs until the initial preprocessing of the data set has been completed.

## 3.3.4 Cross Validation Fold

Cross-validation, a standard evaluation technique, is a systematic way of running repeated percentage splits. Divide a dataset into 10 pieces (“folds”), then hold out each piece in turn for testing and train on the remaining 9 together. This gives 10 evaluation results, which are averaged. In “stratified” cross-validation, when doing the initial division we ensure that each fold contains approximately the correct proportion of the class values. Having done 10-fold cross-validation and computed the evaluation results, Weka invokes the learning algorithm a final (11th) time on the entire dataset to obtain the model that it prints out.

## 3.3.5 Split Percentage:

Evaluation is based on how well it can predict a certain percentage of the data, held out for testing by using the values entered in the ‘%’ field. It's going to make a random split of the dataset. That's because Weka, before it does a run, re-initializes the random number generator. The reason is to make sure that you can get repeatable results.

## 3.3.6 Attribute-Relation File Format (ARFF)

An ARFF [10] (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information.

Header field: The header field describes the name of the attributes, type of relation and their data types that are present in the data file the main difference between them .CSV and .arff file are that the in .CSV files find the values of the attributes just below their name but in .arff files, the name of the attributes are specified separately followed by the data which is present in a separate data field.

Data field: This field contains the data values of the attributes mentioned above in the attribute field these are the values will be used by our model to perform prediction and to determine the amount of accuracy that can be provided in the result of our model. The data present is separated by the comas under the heading of data.

## 3.4 Feature Extraction Model

For feature extraction we have used unigram model and here we have briefly discussed about our feature extraction model.

## 3.4.1 Unigram Model

We have discussed about a unigram model example earlier. To find sequence of probabilities over sequence of terms we can always use the chain rule to decompose the probability of a sequence of events into the probability of each successive event conditioned on earlier events. The chain rule in general:

P(x1,x2,x3,……,xn) = P(x1)P(x2|x1)P(x3|x2,x1)…..P(xn|xn-1,…..x1)

Unigram model, thesimplestformoflanguagemodelsimplythrowsawayallconditioningcontext and estimate each term independently. For such a unigram model:

Puni(x1,x2,x3,……,xn) = P(x1)P(x2)P(x3)……..P(xn)

In this model, the probability to hit each word all depend on its own, so we will only have one- state finite automata as units. Many researchers showed the simplicity and effectiveness of unigram model in their sentiment analysis experiments. That’s why we have used unigram model for our feature extraction.

## 3.4.2 Naïve Bayes

We have discussed Naïve Bayes classifier earlier according to Pang et al [4]. Pang et al [4], Pak and Paroubek [3] have showed the better performance of Naïve Bayes classifier in sentiment analysis as well as text classification. Naïve Bayes is a simple model which works well to perform text classification [18]. For a document ‘d’ and a class ‘c’, Naïve Bayes classifier:

cMAP = arg max P(c|d) …………………………………………..(1)

cϵC

where, MAP is “maximum a posteriori” = most likely class.

= arg max P(d|c)P(c)

[Bayes rule]

cϵC

P(d)

= arg max P(d|c)P(c)

cϵC

because, P(d) plays no rule in selecting c (Pang et al [4]).

= arg max P(x1, x2, x3, … … . . , xn|c)P(c)

cϵC

Document ‘d’ represents as features x1, x2, x3, … … . . , xn

Assuming the feature probabilities P(xi|c) are independent given the class c,

P(x1, x2, x3, … … . . , xn|c) = P(x1|c)P(x2|c)P(x3|c) … … . P(xn|c)…………….(2)

So,

CNB = arg max P(c) ∏xϵX P(x|c)

cϵC

[NB=Naïve Bayes]

## 3.4.3 Multinomial Naïve Bayes

Multinomial Naïve Bayes is a special version of Naïve Bayes that captures word frequency information in documents.

Simply using the frequencies in the data,

P(|)=

Where, v is the vocabulary.

**Chapter 4**

# **Evaluation and Results**

## 4.1 Data Set

We have collected 40001 tweets from website [11] .There were 13 sentiment. Some tweets contain only hashtags, as we want to work only with basic text, we have eliminated approximately 35000 tweets. Some tweets only contain symbols and some tweets contain more than one language, we have eliminated these tweets also. Then we selected total 4751 tweets form there and labeled all the tweets by the help with my partner. While labeling we found some tweets that do not represent any kind of emotion, we have also eliminated these tweets .After that we tried to eliminate the spelling error and on the other hand removed the two alphabet words. Then we have used a data set of total 4751 tweets (happiness contains 785, sadness contains 799, love contains 789, hate contains 798, surprise contains 787 and worry contains 785 data). Eventually, all data have used as a training and testing data.

Basically we choose six categories from 13 categories data sets. We choose happiness, sadness, love, hate, surprise and worry. The purposes of chosen these six categories is that we work in human emotions by using their text. These six categories are the most common and expressing emotions of human beings. We also mainly follow the Abu et al [2] paper for our research. In this paper they have already work in four categories of sentiments. This six categories is really effective for our sentiment analysis. In the purpose of all paper we have chosen these six categories for getting a good sentiments analysis result about human sentiments.

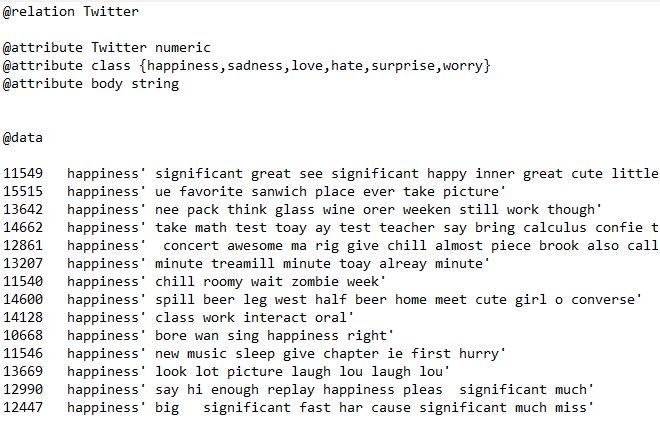
## 4.2 Tools Used For Our Thesis

## 4.2.1 Microsoft Excel and Text Document

We used Microsoft Excel for labeling the tweets. Then we transferred it into .txt file and changed its extension as .arff (Attribute-Relation File Format).

## 4.2.2 Preparing data for classification

We have already mentioned that we used Naïve Bayes Multinomial Text as our classifier. Here we briefly explain the procedure of computation of emotion of a tweet. Firstly, Data is stored in .arff file format specific for WEKA software. There are some significant rules and regulations to make .arff file. Following this all nominal values must be represented between single quote and the obvious things is must have some random values in it’s begin and figure-4.1 represents like this

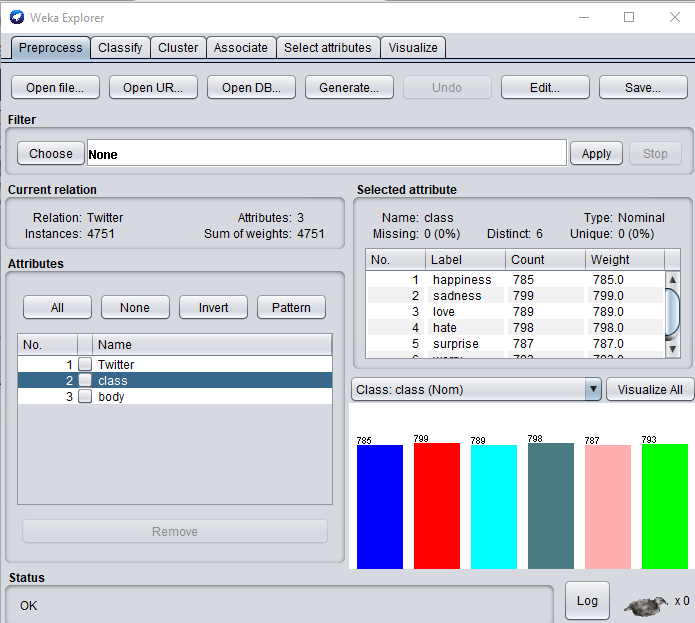
We have already mentioned that we used Naïve Bayes Multinomial Text as our classifier. Here we briefly explain the procedure of computation of emotion of a tweet. Firstly, Data is stored in .arff file format specific for WEKA software. There are some significant rules and ****Regulations to make .arff file. Following this all nominal values must be represented between single quote and the obvious things is must have some random values in its begin and figure-4.1 represents like this:

**Figure 4.1- ARFF file.**

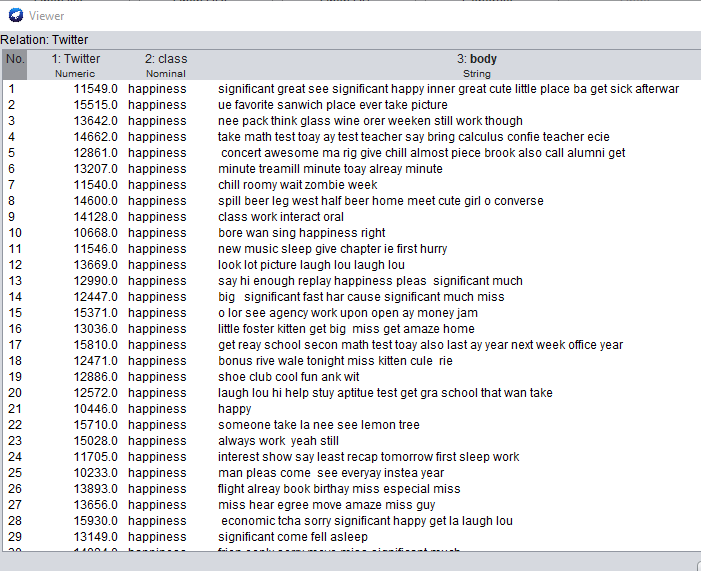
## 4.3 Experimental Results

Here, we can see the attributes (Twitter, class and body) and using the dataset we are going to train Naïve Bayes multinomial Text Model and also apply this model to new data to see to which class it will be assigned.

First of all, figure-4.2 depicts WEKA explorer preprocess tab we need to open our .ARFF data file:

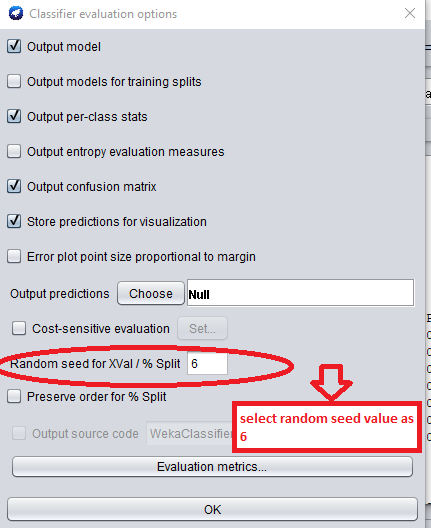
****

**Figure 4.2 -WEKA explorer preprocess tab.**

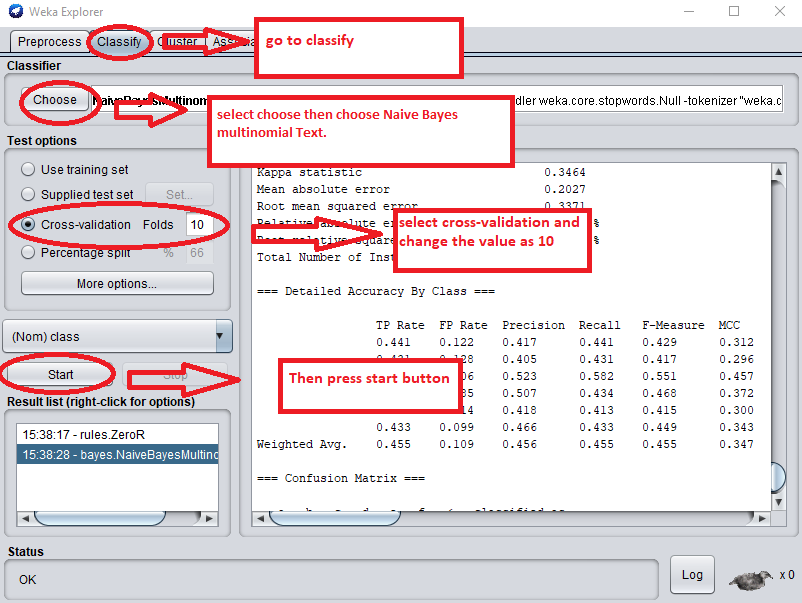
****Here we can see the basic statistics of attributes and user can able to edit runtime ,on the other side observe their data. If we click Edit button, the new Viewer window with data table will be loaded following figure 4.3

**Figure 4.3-Edit tab viewer.**

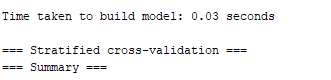
In viewer we can edit data as we like and then we can always save new data set with save button in explorer. We will do when we will create test set with cool and high parameter values. For this we just delete all lines of data except first one and edit values like Figure 4.4.

Evaluation report for given dataset with running results using Naïve Bayes Multinomial Text classification approach. First select classify then select cross validation folds as 10,it divides the total instances by 10 and we can process 475 (total instances 4751) instances at a time.Afterwards go to the option random seed split following figure 4.4 and make it as 4.6 and press ok button.After that select choose then select Naïve Bayes Multinomial Text And our classifier is ready for start and all are following figure 4.5.

**Figure 4.4-classifier evaluation options.**



**Figure 4.5-weka explorer classify layout.**

Evaluation report summery of weka is displaying next illustrates in figure 4.6.

**Figure 4.6-weka building time.**

Screenshot_14.pngThe run time result is displaying that contains 45.5273% (totally 2163 data)as correctly classified instances and rest are the incorrect classified instances in 54.4724%(totally 2588 data). Show it in figure 4.7.

**Figure 4.7-weka accuracy rate.**

Here is the confusion Matrix shown in figure 11 and result shown in table 1 and that contains individually correct classified results:

Happiness correctly classified 44% individually from 785 instances,

Sadness correctly classified 43% individually from 799 instances,

Love correctly classified 58% individually from 789 instances,

Surprise correctly classified 41% individually from 787 instances,

Hate correctly classified 43% individually from 798 instances,

Worry correctly classified 43% individually from 793 instances.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature Extraction Model | Accuracy | | | | | | Average Accuracy |
| Happiness | Sadness | Surprise | Worry | Sad | Love |
| Unigram | 44.08% | 43.05% | 41.30% | 43.23% | 43.05% | 58.17% | 45.53% |

**Table 1- Test accuracy and average accuracy for 6-way classification.**

And all the values are showed in graph in Figure 4.8.

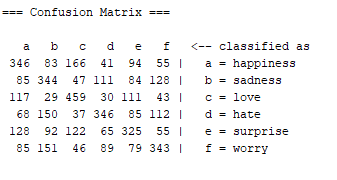
**Figure 4.8- Accuracy of emotion classes for 6-way classification.**

Then using unigram model we go for 2 way classification we assume happiness, love and surprise as positive and sadness, hate and worry as negative. Then we evaluate the results, 57% as positive and 43% as negative. And the average accuracy rate is 49%.That shown in table 2 and the graph depicted in figure 4.9.

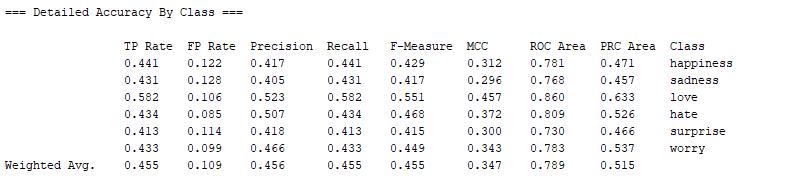
|  |  |  |  |
| --- | --- | --- | --- |
| Feature Extraction Model | Accuracy | | Average Accuracy |
| Positive(Happiness ,Love, Surprise) | Negative(Sadness, Hate, Worry) |  |
| Unigram | 57.62% | 43.22% | 49.72% |

**Table 2-Test accuracy and average accuracy for 2-way classification.**

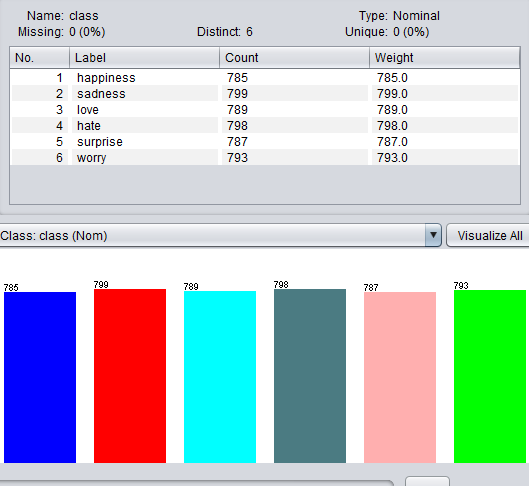
**Figure 4.9 Accuracy of emotion classes for 2-way classification.**

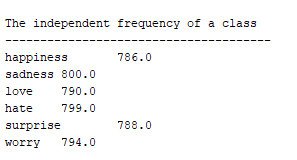
**Figure 4.10- weka confusion matrix**

The figure 4.11 screenshot represents the details accuracy by class and weighted average.

**Figure 4.11-weka detailed Accuracy by class**.

Following screenshot figure 4.12 presenting the visualization of given dataset by attribute of class (Happiness, Sadness, Love, Hate, Surprise and Worry) and show the number of instances contains by the class.

**Figure 4.12-weka visualization of given dataset**.

The screenshot depicts (figure 4.13) the independent frequency of a class.

**Figure 4.13-weka independent frequency of a class.**

## 4.4 Analysis

Table 3 compares the accuracy of our system with the system of Abu et al [2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Extraction | Accuracy of our system | | Accuracy of Abu et al | |
| 6-way classification | 2-way classification | 4-way classification | 5-way classification |
| Unigram | 45.52% | 49.72% | 81% | 66% |

**Table 3: Comparison of accuracy of our system and Abu et al**

Apart from this, Abu et al [2] achieved highest average accuracy of 81% using Unigram in their 4-way classification and for 5-way classification they achieved highest average accuracy of 66% using Unigram with POS tagging. We used 2 way classification (let happiness, surprise and love as positive class and to sharp contrary sadness, hate and worry as negative) and it was just our assumption. Though the average accuracy of 6-way classification was quite low but it’s a huge challenge to work with 6 different classes. Meanwhile, the twitter data doesn’t represent any clear emotion and happiness, love and surprise express quite same and obviously it’s tough to separate them.

**Chapter 5**

# **Conclusions and Future Directions**

## 5.1 Conclusion

Presently a-days microblogging destinations have turned into a piece of our correspondence and public activity. These have a lot on close to home, social, business and governmental issues. So estimation examinations of these microblogging locales have incredible means and scientists as of now have dealt with supposition investigation of these microblogging destinations. We therefore built up a framework for feeling examination with Naïve Bayes classifier for twitter information. Our outcome demonstrates that it can perform better on the off chance that we can prepare more information. Contrasting and related papers, we can see that our outcomes don't vary much from theirs. Our prime concern was to manage just the content segment of a content.

## 5.2 Limitations

There are a few impediments in our work. It would be better in the event that we could do the explore different avenues regarding an enormous arrangement of preparing information. It was extremely tedious for us to gather and mark tweets. In double grouping, in the event that there shows up a word "not" in the tweet, the tweet can be effortlessly ordered into the turnaround notion. Be that as it may, in multiclass characterization, we can't take the choice about the class of the tweet with nullification. For instance, we think about a tweet, "I am not worry", here the tweet plainly does not speak to “anger or disgust”, but we cannot say that the tweet speaks to cheerful. Another issue is to locate the right feeling. On the off chance that we think about a tweet, "I'm not going to class today", here this tweet can speak to “happy or sad”. Just the client who tweeted this knew about his feeling. The feeling of such sort of tweet relies upon the viewpoint.

## 5.3 Future Directions

Since our dataset isn't sufficiently enormous, grouping execution of our framework may increment with a rich dataset. We couldn't work with emoji’s and hashtags, however emoji’s and hashtags speak to feelings. Thus, utilizing emoji’s and hashtag as highlights, precision of our framework may increment. We couldn't analyze our test by utilizing different classifiers. Feeling classes can be expanded. In future, we may work with hashtag and emoticons.

## Bibliography

1. Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011.
2. Abu Zonayed Riyadh and Nasif Alvi "Exploring Human Emotion Via Twitter" analysis on twitter data. In 20th ICCIT in December 22-24 (2017).
3. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In LREc, vol. 10, no. 2010. 2010.
4. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.
5. Pang and Lee, 2008. Online: http://www.nowpublishers.com/article/DownloadSummary/NR-011.
6. Ann Smarty, “5 Tools for Downloading and Analyzing Twitter Data”, February 17,2015. [Online: https://www.entrepreneur.com/article/242830](https://www.entrepreneur.com/article/242830), last accessed: December 12, 2018.
7. A survey on twitter sentiment analysis in 2016. Online: <https://www.researchgate.net/publication/304916478_Like_It_or_Not_A_Survey_of_Twitter_Sentiment_Analysis_Methods>. Last access: 1st December 2018.
8. Weka Data details and formatting detail. Online:https://opensourceforu.com/2017/01/anintroductiontoweka/?fbclid=IwAR0\_Wf0Z5PEF7UfQzULbYho1EoBUkteAXKKO2wjFnEAxamrUwymzao9Zf4. Last Access: December 1st 2018.
9. Weka Explorer by https://opensourceforu.com/2017/01/an-introduction-to-weka/.
10. Attribute-Relation File Format (ARRF). Online: https://www.includehelp.com/ ml-ai/ attribute-relation-file-format.aspx. Last access: 14th December 2018.
11. Data Set Collection of twitter sentiments. Online: https://www.figure-eight.com/ data-foreveryone/?fbclid=IwAR175zMDgdOTenEVNZeifYOqEMocgg7OBqrLJpgJ0Mv8BtmykBS4n6QZLhk. Last access on 15th June 2018.
12. Figure 1.1: Number of monthly active twitter users in millions. Online: https://www.statista.com/ statistics/282087/ number-of-monthly-active-twitter-users. Last access on 03 august in 2018.
13. Raw data. Online: <https://www.brandwatch.com/blog/top-10-free-twitter-analysis-tools/>
14. Raw Data using and analysis tools details. Online: https://www.entrepreneur.com/ article/242430. Last Access: 10th December 2018.
15. Purver, Matthew, and Stuart Battersby. "Experimenting with distant supervision for emotion classification." In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 482-491. Association for Computational Linguistics, 2012.
16. Balabantaray, Rakesh C., Mudasir Mohammad, and Nibha Sharma. "Multi-class twitter emotion classification: A new approach." International Journal of Applied Information Systems 4, no. 1 (2012): 48-53.
17. Waikato environment and analysis(Weka). Online: https://www.cs.waikato.ac.nz/~ml/ publications/1995/ Garner95-WEKA.pdf. Last Access: 14th December 2018.
18. Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
19. Stemming and Lemmatization. Online: [https://nlp.stanford.edu/IRbook/html/htmledition/stemming-and-lemmatization-1.html](https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html). Last access:1sth December 2018.
20. Stemming and Lemmatization. Online: <https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>. Last access: 15th November 2018.
21. “The Twitter glossary”, Online: <https://support.twitter.com/articles/166337>, last accessed: December 10, 2018.
22. Character limits of twitter. Online: [http://time.com/4958311/twitter-280-character-limit/](http://time.com/4958311/twitter-280-character-limit/?fbclid=IwAR3SnD1kGAaRwGB59Ru4VCZLcunuxIZKbt5mDEJJRrZ2npKWM4l4Bpu3jFM) Last access: 1st October 2018.
23. Weka Data details and formatting detail. Online: http://en.wikipedia.org/wiki/Weka­\_(machne\_learning). Last access: December 1st 2018.
24. Alec Go, Richa Bhayani, and Lei Huang, “Twitter Sentiment Classification using Distant Supervision”, CS224N Project Report, Stanford, vol. 1, pp. 12, December, 2009.