Shahjalal University of Science and Technology,

Department of Computer Science and Engineering.

**CSE 408**



Sentiment Analysis - Detect the sentiment of Bangla text

Name: Shoaib Alam Reg. no: 2013331064 Year/Semester: 4/1 Dept: CSE

Name: Fowzia Yesmin Reg. no: 2013331037 Year/Semester: 4/1 Dept: CSE

Supervisor:

Name: Md. Masum Designation: Associate Professor Dept: CSE

09th September , 2018

Recommendation Letter from Thesis Supervisor

This report is for the completion of first part (4/1 semester) of our thesis on “Sentiment Analysis- Detect the sentiment of Bangla text” which is assigned to us by CSE dept. of SUST for the fulfillment of Course No. 408.

Group Members

Shoaib Alam

2013331064

Fowzia Yesmin Munni

2013331037

Md. Masum

Associate Professor

Dept. of Computer Science and Engineering, SUST

**Qualification Form of Bachelor Degree**

Student Name: Shoaib Alam, Fowzia Yesmin Munni,

Thesis Title: Sentiment Analysis – Detect the sentiment of Bangla text.

This is to certify that the thesis submitted by the student named above in March, 2018. It is qualified and approved by the Thesis Examination Committee.

|  |  |  |
| --- | --- | --- |
| **Acting Head of the Dept.** | **Chairman, Exam. Committee** | **Supervisor** |

Dr. Mohammad Reza Selim Dr. MD. Jahirul Islam MD. Masum

Professor & Head Professor Associate Professor

Department of Computer Department of Computer Department of Computer

Science and Engineering. Science and Engineering Science and Engineering

**ABSTRACT**

Sentiment Analysis is an area of important research over the last decade. The basic task in sentiment analysis is classifying a given text. Sentiment analysis is an important type of text analysis process that aims to measure how much positive or negative a text is. Though Bangla is one of the top ten most spoken languages in the world and is spoken by more than 200 million people, it still lacks significant research efforts in the area of Bangla Sentiment Analysis. In the first phase of our thesis work we collected raw data of sentences, which were used as comments by users at Facebook.

In the later phases, we used different machine learning techniques to determine whether the given query is positive or negative.

**Keywords：**

Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbor, Logestic Regression, TF-IDF, Sentiment Analysis

**Acknowledgments**

We wish to express my profound sense of gratitude to our supervisor Assistant Professor Md. Masum for introducing us to this research topic and providing their valuable guidance and unfailing encouragement throughout the course of the work. Besides, we would like to thank our teacher, staffs, juniors, friends, surroundings specially parents for their beloved support, inspiration and motivation.

We are immensely grateful to them for their constant advice and support for successful completion of this work.

TABLE OF CONTENTS

Page

[ILLUSTRATION OF SYMBOLS](#_Toc82246438) 09

[LIST of TABlEs](#_Toc82246438) 10

1 [INTRODUCTION](#_Toc82246440) 10

2 [BACKGROUND STUDY](#_Toc82246456) 11

[2.1 ACTIVITIES ON SENTIMENT ANALYSIS FOR FOR ENGLISH](#_Toc82246454) 12

[2.2 ACTIVITIES ON SENTIMENT ANALYSIS FOR FOR BANGLA](#_Toc82246454) 13

[2.3 SENTIMENT ANALYSIS IN NEWSPAPER & BLOGS](#_Toc82246454) 14

[2.4 SENTIMENT ANALYSIS IN SOCIAL MEDIA](#_Toc82246454) 14

3 [WHY SENTIMENT ANALYSIS](#_Toc82246456) 15

4 [DATA COLLECTION](#_Toc82246456) 16

5 [METHODOLOGY](#_Toc82246448) 18

[5.1 FEATURE SELECTION](#_Toc82246449) 18

5.1.1 COUNT VECTOR……………………………………………………...18

5.1.2 TF-IDF………………………………………………………………….18

5.2 MODEL SELECTION…………………………………………………..19

5.2.1 NAIVE BAYES………………………………………………………..19

5.2.2 DECISION TREE……………………………………………………...19

5.2.3 LOGESTIC REGRESSION……………………………………………20

5.2.4 K-NEAREST NEIGHBOR…………………………………………….20

5.2.5 SUPPORT VECTOR MACHINE……………………………………..21

6 PERFORMANCE ANALYSIS21

6.1 COUNT VECTOR AS FEATURE

6.2 TF-IDF AS FEATURE

7 CONCLUSION 24

[8.1 FUTURE SCOPE](#_Toc82246454) 24

[8.2 CONCLUSION](#_Toc82246454) 24

8 REFERENCES 25

**List of Pictures**

**Page**

|  |  |  |
| --- | --- | --- |
| Picture 1 | Corpus Of the Analyzer | 16 |
| Picture 2 | Count Vector as Feature | 22 |
| Picture 3 | TF-IDF as Feature | 23 |
| Picture 4 | Decision Tree | 20 |
|  |  |  |
|  |  |  |
|  |  |  |

**Illustration of Symbols**

SVM: Support Vector Machine

KNN: K-Nearest Neighbor

DT: Decision Tree

LR: Logistic Regression

NB: Naive Bayes

TF: Term Frequency

## Chapter 1

## INTRODUCTION

Sentiment analysis is the most interesting and newly emerged research topic in the field of Natural Language Processing (NLP). Basically sentiment analysis is an automated system which uses various machine learning techniques on a user’s or author’s opinion to detect the sentiment. A user may express his/her emotion, excitement, confession, feelings, boredom through a text and a system needs to detect it. Now a day’s social networking sites are getting popular exponentially. Most of the business minded people use social networking sites for advertising their product and getting review from user’s or customers. Besides that, people write blogs, advertise movie and many other things. One can easily know how people are reacting about their works through sentiment analysis.

Sentiment analysis can be categorized into two major parts positive and negative. In this respect, a sentiment analysis task can be interpreted as a classification task where each category represents a sentiment. Sentiment analysis for a company with a means to estimate the product acceptance and to determine the strategy to improve the product quality. To analysis a sentiment we need to build a classifier or patterns to classify a specific type of sentiment. If one pattern or classifier to fail to classify one document then it will check the other patterns until it finds a suitable classifier. In this thesis we work on sentiment analysis in order to detect the sentiment of a Bangla text.

Chapter 2

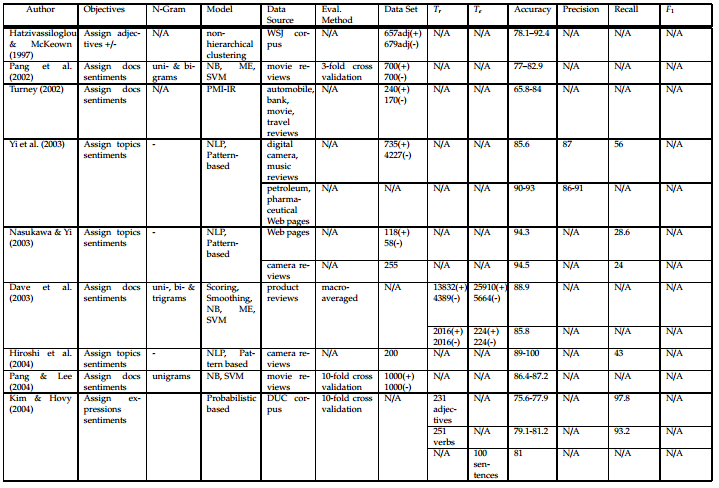
Background study

## 2.1 Activities on Sentiment Analysis for english: [1]

While most researchers focus on assigning sentiments to documents, others focus on more specific tasks: finding the sentiments of words (Hatzivassiloglou & McKeown 1997), subjective expressions (Wilsonet al. 2005; Kim & Hovy 2004), subjective sentences (Pang & Lee 2004) and topics (Yi et al. 2003; Nasukawa & Yi 2003; Hiroshi et al. 2004). These tasks analyze sentiment at a fine-grained level and can be used to improve the effectiveness of a sentiment classification, as shown in Pang & Lee (2004). Instead of carrying out a sentiment classification or an opinion extraction, Choi et al. (2005) focus on extracting the sources of opinions, e.g., the persons or organizations who play a crucial role in influencing other individuals’ opinions. Various data sources have been used, ranging from product reviews, customer feedback, the Document Understanding Conference (DUC) corpus, the Multi-Perspective Question Answering (MPQA) corpus and the Wall Street Journal (WSJ) corpus. To automate sentiment analysis, different approaches have been applied to predict the sentiments of Words, expressions or documents.

These are Natural Language Processing (NLP) and pattern-based (Yiet al. 2003; Nasukawa & Yi 2003; Hiroshi et al. 2004; K¨onig & Brill 2006), machine learning algorithms, such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM) (Joachims 1998),and unsupervised learning (Turney 2002).

Table 2 lists some existing work in this area, and shows different types of objectives along with the associated models used and the experimental results produced.

****

**TABLE 1[1]**

## 2.2 Activities on Sentiment Analysis for Bangla:

Some work has already done for Bangla language. Such as, by using the WorldNet to get the senses of each word according to its parts of speech and SentiWordNet to get the prior valence (i.e. polarity) of each word. We calculate the total positivity, negativity and neutrality of sentence or document with respect to total sense **[3].**

Used SentiWordNet, WorldNet API for lexical analysis and implement support vector machine and maximum entropy model to detect the sentiment of microblog. Besides, they gave preference of emoticons **[4].**

# **2.3 Sentimental analysis in newspaper & Blogs:**

Newspapers and blogs express opinion of news entities (people, places, things) while reporting on recent events. We present a system that assigns scores indicating positive or negative opinion to each distinct entity in the text corpus.

News can be good or bad, but it is seldom neutral. Although full comprehension of natural language text remains well beyond the power of machines, the statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense of how the latest news impacts important entities.

Recent years have brought a significant growth in the volume of research in sentiment analysis, mostly on highly subjective text types (movie or product reviews). The main difference these texts have with news articles is that their target is clearly defined and unique across the text. Following different annotation efforts and the analysis of the issues encountered, we realized that news opinion mining is different from that of other text types. We identified three subtasks that need to be addressed: definition of the target; separation of the good and bad news content from the good and bad sentiment expressed on the target; and analysis of clearly marked opinion that is expressed explicitly, not needing interpretation or the use of world knowledge. Furthermore, we distinguish three different possible views on newspaper articles - author, reader and text, which have to be addressed differently at the time of analysing sentiment.

**2.4** **Sentimental analysis in Social media:**

This thesis presents a comparison of different machine learning techniques applied to the case of sentiment analysis in social media. Social media sites like Facebook and Twitter have become extremely popular since their appearance. Today millions of people share their impressions about the world with their friends and acquaintances in social media. In today’s connected world, users can send messages in any time. However, social media is not only used as a casual tool for messaging and sharing private things and thoughts; it is also used by journalists, politicians and public figures, series of companies and universities who want to be more open to the public, share their thoughts and take an interest in opinion of persons. The active growth of the audience of social media on the Internet led to the formation of these resources as a new source of the people’s mood and opinion. The tracking of citizens’ reactions in social media during crises has attracted an increasing level of interest in the research community .

Social media can be referred to as the ”group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content”, as defined by Kaplan and Haenlein [40]. In recent years in addition to the leaders of **the World Wide Web** such as **Facebook, Google+, LinkedIn and Twitter**, there are new services for different groups of users: social network for students, the network for specific groups of professionals, communities of ethnic minorities, and even a special network for all the world’s drinkers. This extends the scope to very different kinds of research from consumer preferences to psychological characteristics.

Facebook dominates the global social media landscape, claiming 1.366 billion active users in January 2015. Meanwhile, instant messenger services and chat apps continue to grow, with **WhatsApp**, **WeChat**, **Facebook Messenger** and **Viber** all reporting more than 100 million new monthly active users over the past 2014. Instant messenger services and chat apps now account for 3 of the top 5 global social platforms, and 8 instant messenger brands now claim more than 100 million monthly active users. In Twitter, the number of monthly active users is 284 million in 2015. In 2016 the number of monthly active users exceeded 320 million.

The simplest and most common polarity scheme assumes two categories, positive and negative. These two categories constitute the extreme ends of a discrete or continuous scale. This definition covers most voting schemes used in practice, such as

• thumbs up/down (e.g., Facebook, YouTube)

• positive, neutral, negative (e.g., eBay)

• star ratings (e.g., Amazon, IMDb).

## chapter 3

**WHY SENTIMENT ANALYSIS:**

Sentiment Analysis is a very new topic in Bangla Language. No work has been done in that topic so far. Today’s world is based on statistic. So, automated sentiment analyzer can build up a basement to predict a specific product or topic acceptance to others. 24% people in Bangladesh pay income tax is more reliable information than few people in Bangladesh pay income tax. Statistical report or analysis is more reliable and believable to others than assumption. In our thesis, our required destination is to polarize human’s document or statement and built a statistical report.

For an example;

|  |  |  |
| --- | --- | --- |
| Comment | Positive | Negative |
| রুবেল আজ হ্যাপি করে দিল । | 1 | 0 |
| বাংলদেশ আজ ভাল খেলেও ম্যাচটি হেরে গেল। | 0 | 1 |

**Table 2**

So,

Total comment = 2

Positive comment = 1

Negative comment = 1

Percentage of positive comment = 50%

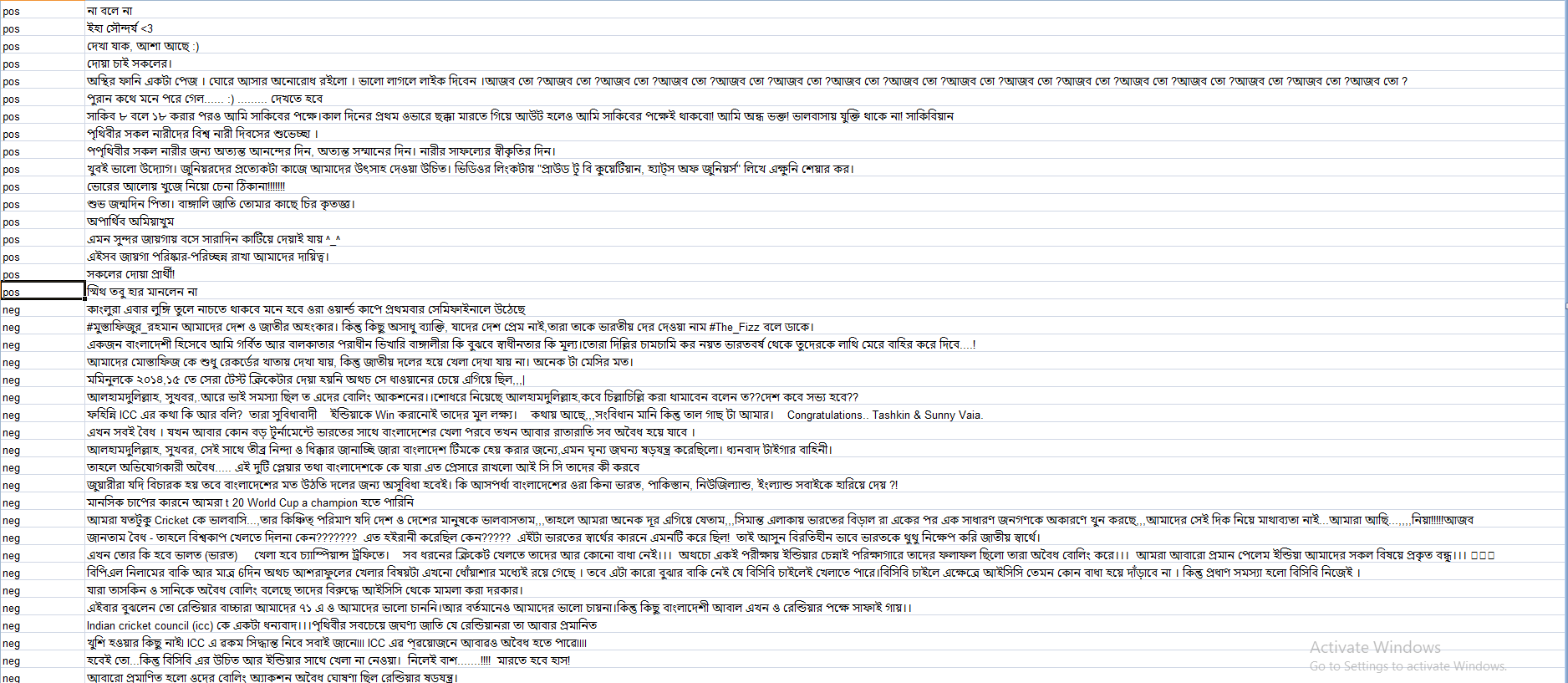
Percentage of Negative comment = 50%

So, we can easily say that 50% people comment positively about today’s match and 50% people comment negatively.

## chapter 4

**DATA COLLECTION**

The size of our initial collected data is more than 1,00,000 sentences. The sentences in the dataset were sourced from Facebook which were posted by users as comments. In our dataset, there are sentences of various types. There are not only Bangla and English sentences, but also, Banglish sentences. The figure below presents a snapshot of our collected dataset.



## 4.1 Sentiment Analysis on social networking site

Two most popular and common social networking sites are Facebook and Twitter. Generally, Facebook is more popular and common than Twitter. Facebook also contain more Bangla comments and review than others. We are working on Bangla language so Facebook is the best option to retrieve data.

Many work has already done to analysis the sentiment of Twitter data. And the data set on Twitter is English. But sentiment analysis on Facebook Bangla data is rare.

**4.2 FACEBOOK USER CLASSIFICATION**

If we analyze Facebook we found that there are 5 types of user. They are picture people, commenters, likers, statusers and the nothings.

**4.3 FACEBOOK COMMENT CLASSIFICATION [2]**

According to **[2]** we can categorize Facebook comments into 14 parts. They are angry, evil, frown, rock-on, shock, smiley, wink, angel, blush, fail, neutral, shades, skeptical, and tongue. For binary classification, we chose to classify into positive and negative sentiment labels. The labels are represented by:

Positive (Smiley, Wink, Tongue, Angel, Shades, Blush, Rockon,)

Negative (Frown, Shock, Skeptical, Evil, Angry, Fail)

**4.4 DATA PROCESSING**

From the initial dataset of 1,00,000 sentences we filtered out the 40,000 Bangla sentences. We tagged the sentences as positive and negative manually, for around 8,000 of such sentences. This is how we processed and prepared our raw data set to create a perfectly stable corpus.

## chapter 5

## Methodology

**5.1 Feature Selection**

**5.1.1 Count Vector**

Text data generally requires special preparation before it can be used for predictive modeling.

For Count Vectorization, the text is parsed to remove words and it is called tokenization. Afterwards, the words need to be encoded as integer or floating point values.

The values are used as input to machine learning algorithms. This process might be called feature extraction or vectorization.

[Count Vectorizer](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html) provides a simple way to both tokenize a collection in text documents as well as build a vocabulary of known words. It is also vital to encode new documents using that vocabulary.

**5.1.2 TF-IDF**

TF-IDF is an information retrieval technique that weighs a term’s frequency (TF) and its inverse document frequency (IDF). Here, every word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF-IDF weight of that term. The TF of a word is the frequency of a word (i.e. number of times it appears) in a document.

The TF-IDF algorithm is generally used to weigh a keyword in any content and assign the importance to that keyword based on the number of times it appears in the document. Moreover, it checks how relevant the keyword is throughout the web, which is referred to as *corpus*.

For a term t in a document d, the weight Wt,d of term t in document d is given by:

Wt,d = TFt,d log (N/DFt)

Where:

* TFt,d is the number of occurrences of t in document d.
* DFt is the number of documents containing the term t.
* N is the total number of documents in the corpus.

**5.2 Model Selection**

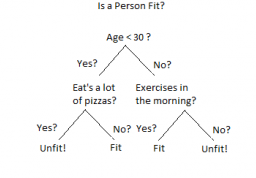
**5.2.1 Naive Bayes**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features.

Though it is based on over-simplified assumptions, Naive Bayes classifiers have worked quite well in many real-world situations. Famous document classification and spam filtering are among the few with exceptional outputs using Naïve Bayes. They require a small amount of training data to estimate the necessary parameters.

**5.2.2 Decision Tree**

Decision Trees are a type of Supervised Machine Learning. Here, the data is continuously split depending on a certain parameter which is previously set. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes whereas the decision nodes are where the data is split.



An example of a decision tree can be explained using above binary tree. If we want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like ‘What’s the age?’, ‘Does he exercise?’, ‘Does he eat a lot of pizzas’? And the leaves, which are outcomes like either ‘fit’, or ‘unfit’. In this case this was a binary classification problem.

**5.2.3 LR**

Logistic regression is named so since the function used at the core of the method is the logistic function. The [logistic function](https://en.wikipedia.org/wiki/Logistic_function), also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.

The coefficients in the logistic regression algorithm must be estimated from the training data and it is done using maximum-likelihood estimation.

[Maximum-likelihood estimation](https://en.wikipedia.org/wiki/Maximum_likelihood) is a common learning algorithm used by a variety of machine learning algorithms, although it does make assumptions about the distribution of the data.

**5.2.4 KNN**

KNN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning" \o "Instance-based learning) where the function is only approximated locally and all computation is deferred until classification. The KNN algorithm is among the simplest of all [machine learning](https://en.wikipedia.org/wiki/Machine_learning" \o "Machine learning) algorithms.

In case of classification and regression, a useful technique to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones can be used. For example, a common weighting scheme consists in giving each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class is known. This can be considered as the training set for the algorithm, in spite of no explicit training step is required.

A peculiarity of the KNN algorithm is that it is sensitive to the local structure of the data.

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the [feature vectors](https://en.wikipedia.org/wiki/Feature_vector" \o "Feature vector) and class labels of the training samples.

**5.2.4 SVM**

Support vector machines are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). Considering a set of training examples, each belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other. This is why it is a non-[probabilistic](https://en.wikipedia.org/wiki/Probabilistic_classification) [binary](https://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

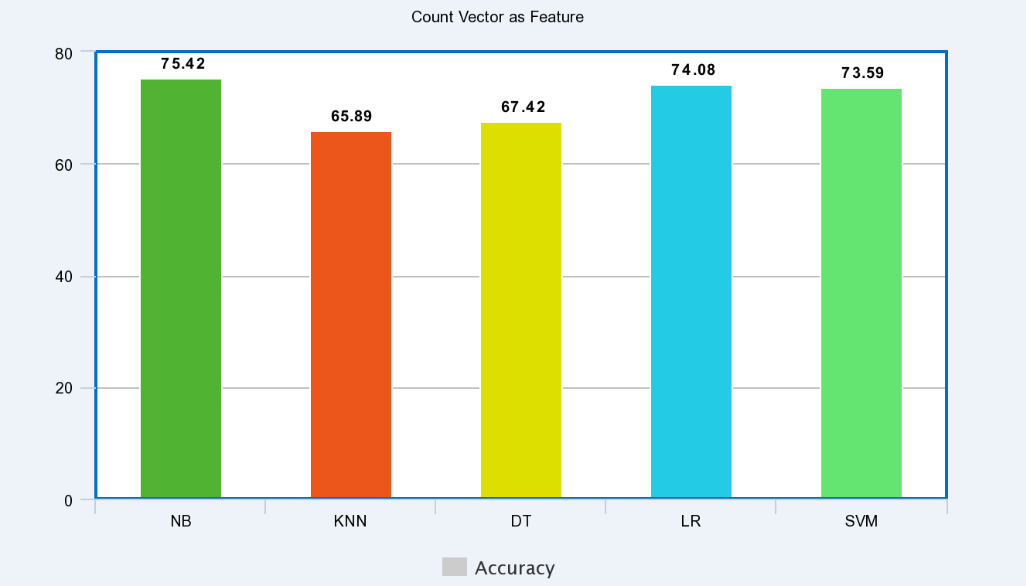
## chapter 6

**PERFORMANCE ANALYSIS**

Implementing the machine learning classifiers using our dataset, we have come across varied performance results. The following sections discusses the results of performance analysis for different features.

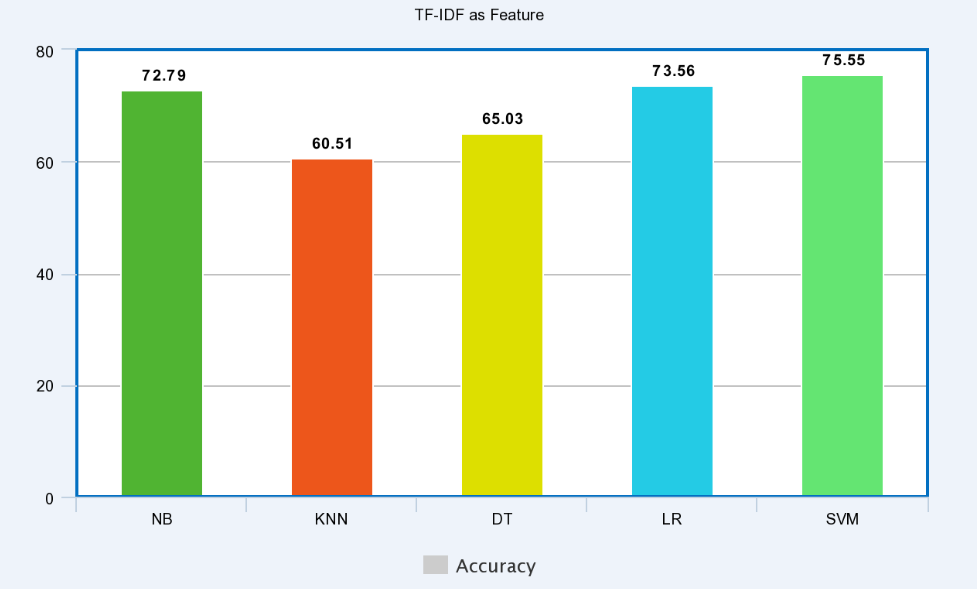
**6.1 Count Vector as Feature:**

Among all the applied classifiers, taking count vector as the selected feature, the Naïve Bayes classifier yielded the best result. Generally, the Naïve Bayes classifier produces better results when we select count vector during the feature selection. The performance of Logestic Regression and Support Vector Machine classifiers came very close to matching that of Naïve Bayes. But, K-Nearest Neighbor and Decision Tree classifiers produced performance which were not such satisfactory. The figure below illustrates further the performances of the classifiers.



**6.2 TF-IDF as Feature:**

When we selected TF-IDF as the feature there were significant changes in the performance of the applied classifiers. Accuracy rates produced by all the classifiers decreased, except for the Support Vector Machine classifier. The SVM classifier performed the best among the classifiers in this case, where as Logestic Regression and Naïve Bayes classifiers produced lower accuracy rates but close to the best result. K-Nearest Neighbor and Decision Tree classifiers produced significantly lower accuracy rates compared to the other classifiers applied. The following figure illustrates the performance of the classifiers.



Therefore, the concluding results show that the Naïve Bayes classifier performed better than any other classifier applied and produced an accuracy of 75.42% when count vector was taken as the feature. But, the Support Vector Machine classifier outperformed all other classifiers with an accuracy rate of 75.55% when TF-IDF was taken as the feature. It is to be noted that we used linear kernel as parameter while implementing the Support Vector Machine classifier.

## chapter 7

**CONCLUSION**

**7.1 Future Scope**

Sentiment analysis is the most interesting and newly emerged research topic. It will open a new door for the writers, bloggers, and businessman. One can easily know the percentage of product acceptance and make their strategy to improve the product quality. Our model can also be used to accurately predict stock market and other business strategical decisions. We think different type of Machine learning model will enforce the result to a significant level.

**7.2 Conclusion**

In this thesis report, we have designed Sentiment Analyzer which detects the sentiment of a text or statement. We used several supervised machine learning methods. It gives us approximately satisfactory accuracy. The approach presented here is flexible and suggests promising avenues for further research. So we conclude our hypothesis as one of the most flexible and accurate implemented model in the field of sentiment analysis.

## chapter 8

**REFERENCES**

**[1]**Rudy Prabowo, Mike Thelwall Sentiment Analysis: A Combined Approach. School of Computing and Information Technology University of Wolverhampton

**[2]**Julie Kane Ahkter (kanej), Steven Soria (ssoriajr) Sentiment Analysis: Facebook Status Messages Final Project CS224N

**[3]** K. M. Azharul Hasan Mosiur Rahman, Badiuzzaman Sentiment Detection from Bangla Text using Contextual Valency Analysis

**[4]** Shaika Chowdhury Wasifa Chowdhury sentiment Analysis for Bangla Microblog Posts.**[2]**Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). The general inquirer: a computer approach to content analysis. The MIT Press. Swan, R. & Allan, J. (2000). Automatic generation of overview timelines.

**[3]**Thelwall, M. (2000). Web impact factors and search engine coverage. Journal of Documentation, 56(2), 185–189.

**[4]**Thelwall, M. (2008). Fk yea I swear: Cursing and gender in a corpus of MySpace pages. Corpora, 3(1), 83–107.

**[5]**Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40 annual meeting of the Association for Computational Linguistics (ACL), July 6–12, 2002 (pp. 417–424). Philadelphia, PA, USA.

**[6]** Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau Sentiment Analysis of Twitter Data Department of Computer Science Columbia University New York, NY 10027 USA

**[7]** Sabir Ismail, Lecturer Shahjalal University of Science and Technology, Sylhet.

**[8]** [www.facebook.com](http://www.facebook.com) for data.

**[9]** [www.wikipedia.com](http://www.wikipedia.com) for data.

**[10]** Merialdo. 1994. Tagging English text with a probabilistic model. Computational Linguistics, 20(2):155-171

**[11]** Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.