Classification of Sentiment Analysis on Tweets using Machine Learning Techniques

Shivaraju Kethavath

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India

May 2015

Classification of Sentiment Analysis on Tweets using Machine Learning Techniques

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Technology, Dual Degree

in

Computer Science and Engineering

(Specialization: Computer Science)

by

Shivaraju Kethavath

(Roll- 710CS1131)

Under the supervision of

Prof. Sanjay Kumar Jena

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela, Odisha, 769 008, India

May 2015

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.

Certificate

This is to certify that the work in the thesis entitled ***Classification of Sentiment Analysis on Tweets using Machine Learning Techniques*** by ***Shivaraju Kethavath*** is a record of an original research work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Master of Technology, Dual Degree with the specialization of Computer Science in the department of Computer Science and Engineering, National Institute of Technology Rourkela. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

.

Place: NIT Rourkela
Date: May 28, 2015
.

Prof. Sanjay Kumar Jena
Professor, CSE Department
NIT Rourkela, Odisha

Acknowledgment

I am grateful to numerous local and global peers who have contributed towards shaping this thesis. At the outset, I would like to express my sincere thanks to Prof. Sanjay Kumar Jena for his advice during my thesis work. As my supervisor, he has constantly encouraged me to remain focused on achieving my goal. His observations and comments helped me to establish the overall direction to the research and to move forward with investigation in depth. He has helped me greatly and been a source of knowledge.

I am very much indebted to Prof. Santanu Ku. Rath, Head-CSE, for his continuous encouragement and support. He is always ready to help with a smile. I am also thankful to all the professors at the department for their support.

I would like to thank Mr. Jitendra Kumar Rout for his encouragement and support. His help can never be penned with words.

I would like to thank all my friends and lab-mates for their encouragement and understanding. Their help can never be penned with words.

I must acknowledge the academic resources that I have got from NIT Rourkela. I would like to thank administrative and technical staff members of the Department who have been kind enough to advise and help in their respective roles.

Last, but not the least, I would like to dedicate this thesis to my father and mother, for their love, patience, and understanding.

Shivaraju Kethavath

Roll-710CS1131

Declaration

I Shivaraju Kethavath of Computer Science and Engineering with Roll no 710CS1131 hereby declare that the project submitted by me is solely of my work and is not copied from any other source where ever may available and It has not been previously submitted for any academic degree. I had verified my thesis report through Turnitin software for plagiarism. All sources of quoted information have been acknowledged by means of appropriate references.

If in future my work was found to be plagiarized from any other persons work, then in that situation I alone will be responsible for it.

Date: May 28, 2015 **Shivaraju Kethavath**

NIT Rourkela

Abstract

Growth in social media has huge of amount of data which includes reviews about products ,blogs which discuss on the peoples opinion .We can learn sentiment analysis in web mining, data mining ,it is an application of Natural Language Processing. Due to growth in social media all the fortune companies are working on Opinion mining. The basic goal of Sentiment analysis is to ensure the sentence either as positive emotion or negative emotion. Sentiment analysis extracts the sentiments in the form various discussions, forums, blogs. Importance of social media leads in growth of sentiment analysis. For an organization, it wants to know about the peoples opinions on products which it had been released and it conducts surveys of products and opinion polls. Consumers also used to make research on products and price of product by using sentiment analysis. Marketers used to make research about company and products by effective utilization of sentiment analysis.

This thesis contributes to classification of tweets in to either positive or negative using Machine learning techniques such as Nave Bayes classifier, Multinomial Nave Bayes algorithm, or Gaussian Nave Bayes, SVM Classifier, and Decision Tree .Comparative tabulation of performance of above mentioned classifiers is created to critically analyze the sentiment of tweets.

Key words: Opinion Mining, Sentiment Analysis, Natural Language Processing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Natural language processing:

The branch of computer science that researches on the development of systems that can communicate with humans in everyday language is called as Natural Language Processing [7]. In theory, it deals with the range of techniques that compute, analyze and represent naturally happening texts at multi-level analysis of languages for the purpose to make the machine process like human language for different disciplines and applications. NLP algorithms depend highly on machine learning with the majority being statistical. Older implementation of language-processing tasks normally required hard coding of big set of rules. By using machine learning, we can use normal learning algorithms usually in statistical inference, to learn rules by analyzing large corpora of real-world examples. A corpus is a set of documents that were annotated by hand with the correct values to be learned. Types of Natural Language Processing are:

1. **Morphological processing:**

   In this state of language processing, strings are broken into sets of tokens corresponding to the words, sub-words and punctuation forms. Normally, a modification occurs not only by adding prefixes or postfixes but may be by other changes.

2. **Syntax and semantic analysis:**

   A processor that carries out different functions primarily based on syntax and semantic analysis. There are two uses of syntax analysis. One is to check

if a sentence is well-formed and the other is to break it into a structure that gives syntactic relation between them. The same can be achieved by a parser using a dictionary of word definitions and a set of syntax rules. A simple lexicon has the syntactic category of every word, the rules are described by the grammar which signify how they can unite phrases of various types.

3. **Pragmatic analysis:**

   Interpreting the results of semantic analysis considering from the point of view of a specific situation is called pragmatic analysis.

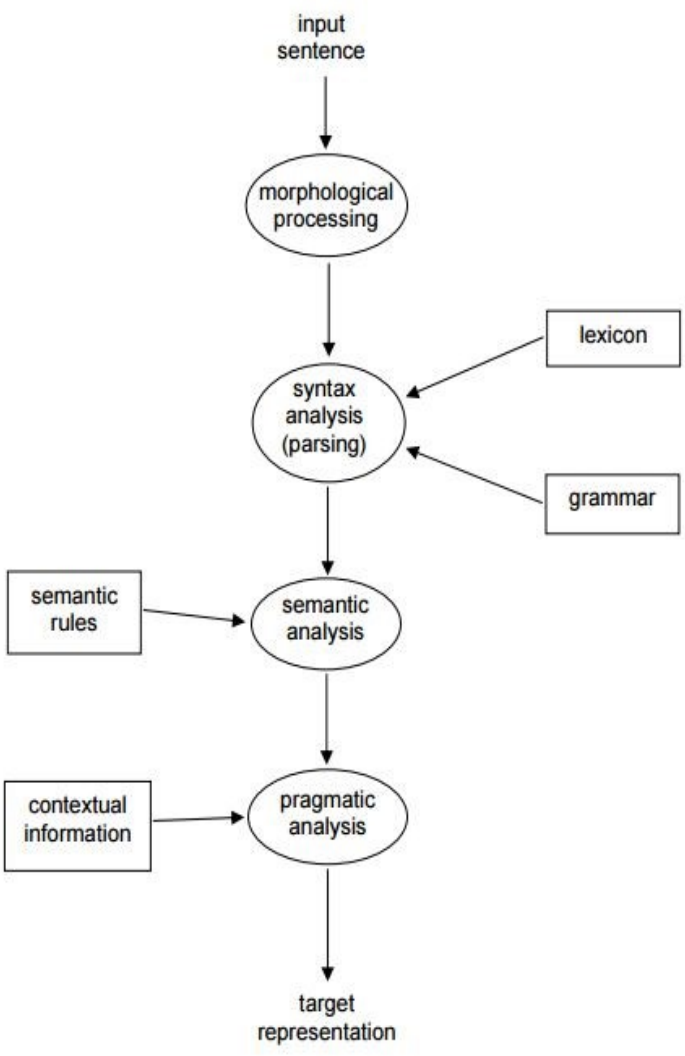

Figure 1.1: Steps of Natural Language Processing

1.2 Sentimental Analysis:

Sentimental Analysis [11] is a method of opinion mining to extract information about peoples views, opinions, sentiments towards an everyday happening things. And an each individual have a different opinions on same topic. The sentiment analysis task is technically more challenging but practically more useful. For example, Businessmen always want to know about the public opinion regarding that products and feedback from different customers. The customers also wants to know the rating of that product which has given by other customers who had purchased earlier, and marketers also prefer sentiment analysis because they wanted to know the targeted customers.

With the major development in social networking (i.e., Facebook, Twitter, LinkedIn, Stumble upon etc.,) on the Web, individuals and large associations are concentrating on publics opinion for their decision making. The task of mining opinion information on web sites is not easy; one because of vast number of websites currently present and still populating and second because of lack of standardized methodology to do the same. Moreover, the text corpora present on websites constitute both useless and useful data that will be required for our analysis. There is always a thin line between these kinds of data which always add unnecessary overhead in analysis. The normal human reader will experience issues distinguishing relevant sites and also summarizing information and opinions in them.

Additionally, it is likewise realized that human analysis of content data is liable to significant preferences, e.g., peoples regularly give more priority to opinions that are reliable with their own preferences. There are other factors as well like human mental capacity and physical limitation that make humans inept to analyze large amount of data. Thus an automated opinion mining is required which will eventually help humans in sentiment analysis.

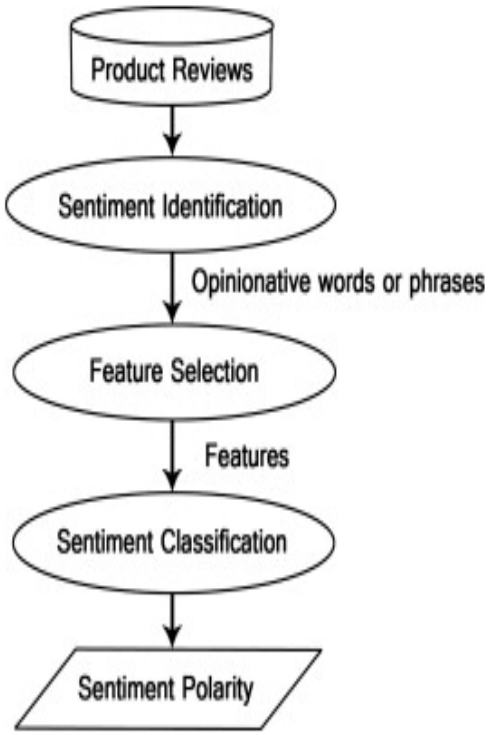

Figure 1.2: Classification of Sentiment Analysis

1.3 Problem Statement

Given a set of tweets containing multiple features and varied opinions, the objective is to extract expressions of opinion describing a target feature and classify it as positive or negative.

1.4 Motivation

Sentiment analysis and opinion minion [14] is open research field with manifold real life applications. Blogs, Forum, Twitter, Facebook and other resources on internet are put to use by humans for expressing their opinions. The social media has bought the people around the world closer; communication is one click away. Before social media there was expensive short messaging service (SMS) provided by telecommunication companies with domestic and international charges. Today the short messaging has evolved from just sending messages to single person to sending messages to multiple people at cheapest price. This service is provided

by many websites but Twitter was the one which pioneered it. Today twitter has hundreds of millions users who post nearly half a billions tweets every day i.e. approximately thousands of tweets for every second. Tweets are not only posted in English language but also in different local languages of the world. These data are precious to business intelligence where the company wants to know "why isnt consumer buying our laptops? ", "why the competitors products are outselling our products". Thus a concrete system to process above mentioned queries is the need of the hour.

The Consumers Perspective:

The feedback of sentiment analysis help consumer to make choices. Previously, people would ask for the opinion of close friends, relatives to review a product. But now with the presence of social media people expression internet [8]. This information is useful for people who are seeking for reviews about a particular product which will help them at taking at making a decision to buy a product. These kind of decisions are generally binary in nature i.e., either the consumer decide to buy it or he doesnt .User cannot make this decision by himself as he would have to go through large volume of data. Hence, this process should be automated with use of technology which will generate good and bad reviews, ultimately helping user in taking decision.

The Producers Perspective:

The decisions of consumers go hand in hand with the decisions of producers. While the consumers are busy sharing their opinions online the producers are monitoring their behavior to take business decisions. This scenario create a multi coupled system of consumers and producers where the nature of consumers expenditure influence others consumers in making decision to buy products and influence the producer to sell the product. The producer learns the trend in sales

of products and services through opinion mining and reshapes its future business plan. Products and services manifest their impression and usefulness through a series of decisions by consumer and producer.

1.5 Objective:

Classify every tweet in either as positive sentiment or negative sentiment using different Machine Learning techniques and check which classifier performs the best.

1.6 Thesis Organization:

Chapter 1: In this chapter we talk about the jargon we are going to use and the methods used in natural language processing after that we discuss about research done on opinion mining.

Chapter 2: Sentiment analysis classification and preprocessing of twitter data.

Chapter 3: We use large number of training datasets of tweets to find polarity by using machine learning techniques.

Chapter 4: In this chapter, we process the twitter dataset with feature of unigram and bigram and test them with substantiated results that have been found on twitter dataset. Finally we get results about which classifier is the best.

Chapter 5: We conclude the thesis and propose our future work based on this.

Chapter 2

Literature Survey

2.1 Sentiment Analysis classification:

1. **Document-level of sentiment analysis:**

   Opinions are the expressions, sentiments or notions towards a component or an event. Numerous data in net or gatherings permit individuals to express their assessment as surveys and remarks. At the point when opinions are communicated as surveys, rather than a straightforward Positive or Negative, recognizing the genuine opinions would require a subjective examination of the words utilized as a part of the survey.

   Opinions on a product will be not the same as one person to other person. There are only two types of classes either Positive or negative .A legitimate case: A thing review: "I brought a new IPhone two days back. It is a good phone. The touch screen is fast. The voice clarity is better. I essentially love this phone". The words which have been used to be utilized. The target sentiments are measured utilizing the star or review framework, where 4 or 5 stars are sure and 1 or 2 stars are negative.

2. **Sentence-level of sentiment analysis:**

   This method use to give useful data when we search because the polarity of sentence will made perfect. In this level of sentiment analysis go through those sentences which contain opinions and gives reviews as though it is negative or positive.

3. **Aspect based sentiment analysis:**

   Document level and sentence level sentiment analysis functions admirably when they allude to a single element. Then again, a great part of the time people talk about components that have various points of view or qualities. They will similarly have unmistakable sentiments about distinctive items [16]. It frequently happens in thing review and dialog social affairs. A valid example: "I am a Nokia phone huge other. I like the look of the phone. The screen is huge and clear. The camera is extraordinary. Nevertheless, there is couple of downsides also; the battery life is not up-to the engraving and access to Whatsapp is troublesome. "Requesting the positive and negatives of this review hides the vital information about the thing. In this way, the Aspect based sentiment analysis focuses on the acknowledgement of all reports within a given record and points to which the feelings.

4. **Comparative sentiment analysis:**

   The users utilization to express diverse emotions on items or brands. Either it may be same item or brand. The objective of comparative sentiment is to discover supposition of relative sort sentence.

5. **Sentiment lexicon acquisition:** Sentimental analysis is a process which utilizes data to find opinions and expressions of that data. In Sentiment analysis, we will see two kinds of classes positive and negative [12]. Given a statement: "Auto X is superior to anything auto Y". This statement doesnt show which class is that statement falls in. Likewise, these sorts of sentences/documents are analyzed using three systems: Manual methodology, dictionary based approach and corpus-based approach.

   - **Manual Methodology:** Its totally time taking process because we cant retrieval the data is in positive or negative.
   - **Dictionary based approach:** This approach utilizes sentiwordnet to find the polarity of that sentence by POS tagging.
   - **Corpus-based approach:** This process is a domain-specific senti-

ment lexicon it is used to carry out the analysis. These are the diverse approaches to investigate customer's opinions and also to predict the market value of a particular organization.

2.2 POS Tagging

POS Tagging is extremely valuable in Opinion Mining procedure [5]. When we have to examine an document or a sentence first we need to concentrate the subjective data from the record or that specific sentence. POS Tagging helps us to find parts of speech of that word. Subsequent to extricating these words we can perform different activities on these and we can reach a conclusion.POS Tagging is done by utilizing the HMM model which used to tokenize and Tag the words furthermore for naming elements.

The word in the content (or the sentence) is tagged utilizing a POS-tagger with the goal that it appoints a name to every word, permitting the machine to do something with it. It looks something like this:

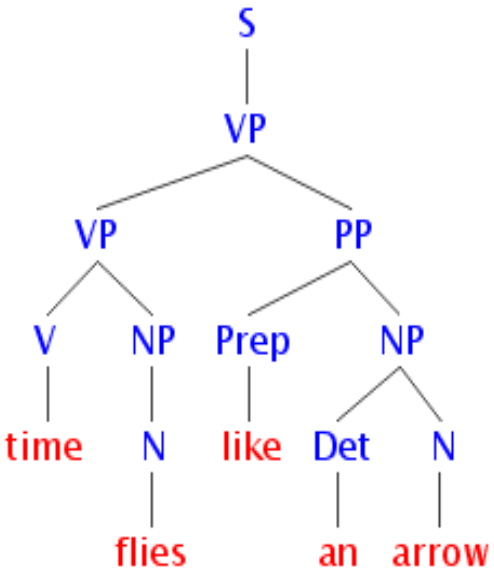

Figure 2.1: POS Tagging

We take sentiment orientation (SO) of the examples are extracted. For instance we may have taken: Amazing + Phone which is:

[JJ] + [NN] (or descriptive word took after by thing in human)

The inverse may be "Repulsive" for instance. In this stage, the machines try to arrange the words on an emotive scale (in a manner of speaking).

The normal Sentiment introduction of the considerable number of expressions we assembled is processed. This permits the machine to say something like: "By and large individuals like the new iPhone" They prescribe it or "For the most part individuals hate the new iPhone" They don't suggest it.

2.3 Twitter Dataset:

Micro blogging stages are utilized by distinctive individuals to express their opinions about diverse themes, accordingly it is an important wellspring of individuals' opinions.

- Twitter contains a gigantic number of content posts and it develops consistently. The gathered corpus can be self-assertively huge.
- Twitter's crowd fluctuates from normal clients to superstars, organization delegates, lawmakers, and even nation presidents. Along these lines, it is conceivable to gather content posts of clients from diverse social and intrigues bunch.
- Twitter's crowd is spoken to by clients from numerous nations. Despite the fact that clients from U.S. are winning, it is conceivable to gather information in diverse languages.

Writings containing positive opinions, for example, satisfaction, beguilement or delight. Target messages that just express a certainty or don't express any feelings we perform an linguistic analysis of our corpus and we demonstrate to construct a sentiment classifier that uses the gathered corpus as training data [6].

Given a message, order whether the message is of positive, negative, or impartial sentiment. For messages passing on both a positive and negative sentiment, whichever is the more grounded opinion ought to be chosen.

2.3.1 Sentence Weightage:

1. If a tweet consists of more than one sentence, we give more weightage to sentences coming afterwards.
2. This is due to the tendency of most tweets to be conclusive in nature.
3. When testing it on small set of tweets, it improved accuracy by around 2.5%.

2.3.2 Hashtag

1. We plan to use the hash tags to get idea about the tweets
2. The hashtags are like this: #IndiabeatAus #FinallySuccessful and so on
3. These hashtags would be structured though not complete sentences
4. So, we would need to parse these tweets before processing
5. Hashtags like #happy, #good, #unhappy, etc give sufficient information about the polarity of the tweets.

2.3.3 Abbreviations and Redundant/Repeated letters

1. Due to the easygoing way of Twitter dialect, a few words (as a rule conclusion words) are incorrectly spelled or frequently over underscored because of which the classifier may not quality polarity of this word (eg. looooooooove) to the genuine word (eg.love) during training.
2. In words containing more than 3 occurrences of the same letter together, these occurrences are supplanted with 2 occasions of the letter. eg. haaaaaaaappy would be changed as haappy , goooooooood would be changed as great.
3. Created a rundown of regular and most prominent shortenings of most generally utilized.

2.4 Methodology

1. Extract all the features from the given review In the absence of any prior information about the domain of the review (in the form of untagged or tagged data belonging to that domain), this will give a list of potential features in that review which needs to be pruned to obtain the exact features. Consider the review, I wonder how can any people like Max, given its pathetic battery life, even though its multimedia features are not that bad. Here, multimedia features and battery life are the exact features pertaining to the mobile domain. But without any prior domain information, we can use an approximate method to obtain a list of potential features that may include other noisy features as well, example people. So this list needs to be pruned to remove the noise and obtain the exact set of features.

2. Extract opinion words referring to the target feature The opinion words are not only Adjectives like hate, love but also consist of other POS categories like Nouns (terrorism), Verbs (terrify) and Adverbs (gratefully). A nave method, like extracting the opinion words closest to the target feature, does not work so well when the sentence has multiple features and distributed emotions (as we will see later). In the example above, pathetic and not bad are the opinion expressions referring to battery life and multimedia features respectively.

3. Classify the extracted opinion words as positive or negative.

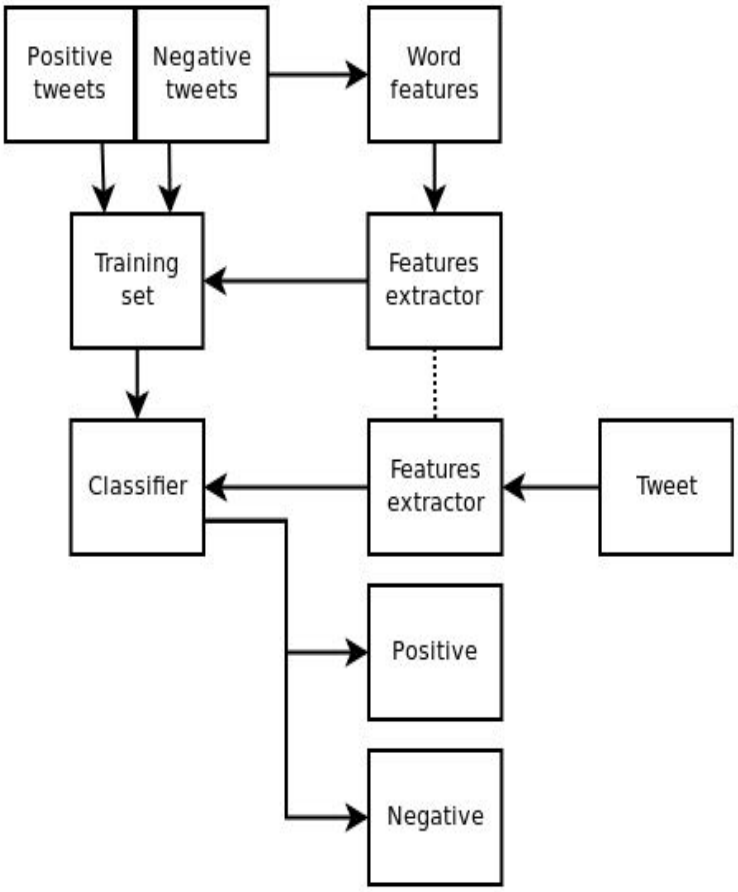

Figure 2.2: Steps to extract polarity of tweets

2.5 Sentiwordnet

Sentiwordnet is a lexical asset of opinion mining .Sentiwordnet allots a synset of word net in three scores: positive, negative, neutral. It extracts the parts of speech of that words that going hand in hand with we see the separated words using POS Tagging from substance documents containing customer reviews. In the following figure the outline between number of reviews and number of words removed for diverse number of reviews.

.

Figure 2.3: Sentiment of a word

Chapter 3

Proposed Work

3.1 Nave Bayes classifier

The basic dependency of Nave Bayes classifier [1] is that predictors are independent. In case of very big data sets it is very easy to create a Nave Bayes model without any iterative estimation of Parameters, which also performs very well when compare to other advanced classification methods.
Bayes theorem can be computed by doing posterior probability, P(c—x), from P(c), P(x), and P(x—c). Nave Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) does not depend on the values of remaining predictors. This is also known as conditional independence.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) = P(x_1|c) * P(x_2|c) * ..... * P(x_n|c) * P(c)$$

1. $P(c|x)$ is the posterior probability .
2. P(c) is the prior probability.
3. $P(x|c)$ is the likelihood.
4. P(x) is the prior probability of predictor.

Confusion Matrix:

	Positive	Negative
Positive	1577	423
Negative	329	1671

.

Accuracy : 81.21%

3.2 Multinomial Nave Bayes algorithm

Multinomial Nave Bayes [1] theorem is similar to nave Bayes algorithm except for the part that it is implemented for multinomial distributed data, and is a classic nave Bayes which is mainly used for text classification. While the simple one models a document if some particular words are present or not, multinomial one models based on the word counts and adjusts the underlying calculations to deal with it. In this model, the attributes are positions of the text and words are considered as values. Before doing this, we assume that the classification does not depend on the position of the words and the similar parameters are used for each position.

If in the training data, the feature and class value does not occur together, then the probability estimate based on frequency will be zero. This results in problem of wiping out information in remaining probabilities when they are multiples. So, a new term called pseudo count is introduced which is a small-sample correction in estimates so that, none of the probabilities is set to zero. This method to regularize Bayes is called as Laplace smoothing if pseudo count is one and lid stone

smoothing otherwise. Normal assumptions while we deal with continue values that are with each class are distributed according to a Gaussian distribution.

The probability distribution of some value given a class $p(x = v|c)$, can be calculated by using v in the equation for a Normal distribution where μ_c and σ_c^2 are parameters.. That is,

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Confusion Matrix:

	Positive	Negative
Positive	1814	186
Negative	215	1785

.

Accuracy : 89.75%

3.3 Decision Tree

The main theme of decision tree [3] is used to characterization technique that produces tree structure where every node signifies a test on a trait worth and every branch speaks to a result of the test. The tree leaves speak to the classes. The figure demonstrates the choice tree assessed from our Training dataset utilized as a part of the project. It shows the connections found in the Training dataset. This method is quick unless the preparation information is substantial. It doesn't give any suspicions about the probability distribution of that particular data. The

procedure of building the tree is called induction.

Building a Decision Tree:

The decision tree algorithm is a top-down greedy algorithm which means to manufacture a tree that has leaves as homogenous as could reasonably be expected. The real step in the algorithm is to keep isolating leaves that are not homogeneous into leaves that are as homogeneous as could be expected under the circumstances until no further division is conceivable. The algorithm is:

1. In the event that a percentage of the traits are ceaseless esteemed, they ought to be discretized into classifications.
2. If the training dataset occurs in same class, then the event will stop.
3. Part the following node by selecting an attribute from the independent attributes that best partitions the articles in the node into subsets and make decision tree.
4. Part the node as indicated by the selected attribute chosen in the step 3. Stop if any of the accompanying conditions meets, generally proceed with step 3:

Confusion Matrix:

	Positive	Negative
Positive	1654	346
Negative	305	1695

.

Accuracy : 80.08%

3.4 SVM Classifier

SVM [2]classifies the information into two different classes which is separated by hyper plane and the classes are either may be positive or negative. Classifying the information is a machine learning task. If a given data is in one of the two classes then we need to decide which class will be fitted by new data point. SVM with largest margin is better one . The points which are closest to hyper plane is support vector.

This figure shows linear classification, with + indicating data points of type 1,

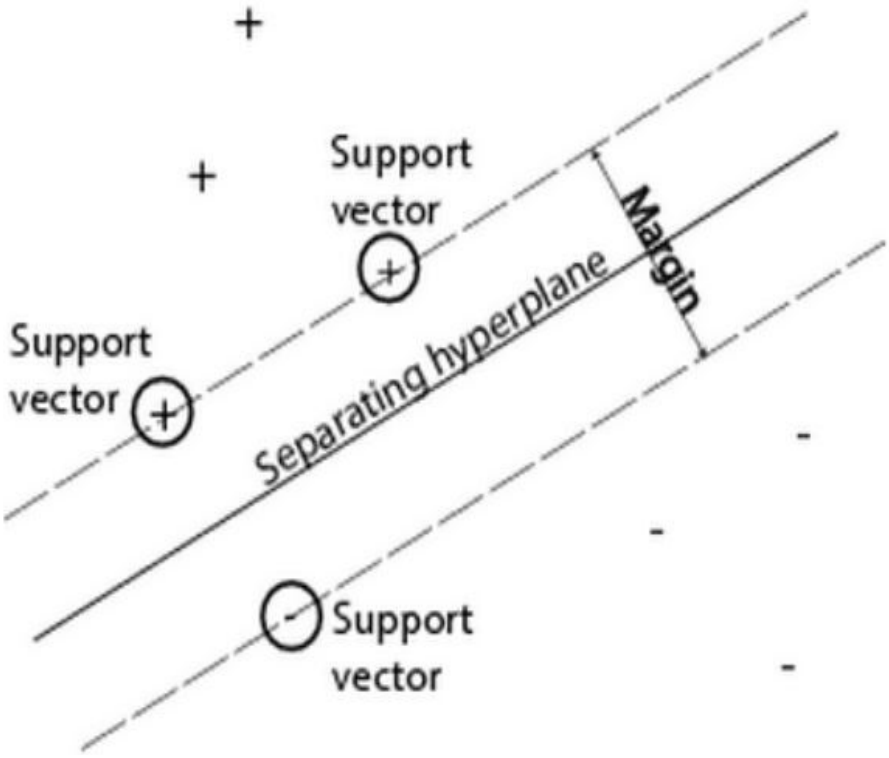

Figure 3.1: SVM Hyperplane

and indicating data points of type 0. The datasets that we have used cannot be classified using linear classifier. Nonlinear classifier with Gaussian kernel is used.

Confusion Matrix:

	Positive	Negative
Positive	1893	107
Negative	205	1885

.

Accuracy : 94.45%

Chapter 4

Results

	Naive Bayes Classifier							
	Unigram		Unigram + POS		Bigram		Bigram + POS	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Positive	1577	423	1528	472	1464	536	1408	592
Negative	329	1671	390	1610	477	1523	507	1493
Accuracy	81.21		78.45		74.67		72.25	

Confusion Matrix constructed by Naive Bayes Classifier

The trained data set which have been classified by Naive bayes classifier and the test corpora with features of unigram and bigram by performing POS tagging it shows a better result at unigram model with Naive Bayes classifier .Naive Bayes Classifier is basic approach of Machine Learning Techniques.

	Multinomial Naive Bayes Classifier							
	Unigram		Unigram + POS		Bigram		Bigram + POS	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Positive	1867	123	1814	186	1606	394	1586	414
Negative	197	1803	305	1695	477	1523	522	1478
Accuracy	91.75		89.97		78.22		76.6	

Table 4.1: Confusion Matrix constructed by Multinomial Naive Bayes Classifier

Multinomial Nave Bayes theorem is similar to nave Bayes algorithm except for the part that it is implemented for multinomial distributed Naive Bayes Classifier is classified by a trained data set .The features used here is uni-gram and bi-gram. By performing classification it Shows better result than Naive Bayes classifier and uni-gram feature shows a better result in it.

	Decision Tree Classifier							
	Unigram		Unigram + POS		Bigram		Bigram + POS	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Positive	1654	346	1474	526	1424	576	1378	622
Negative	305	1695	417	1583	502	1498	547	1453
Accuracy	80.08		76.42		73.05		70.75	

Table 4.2: Confusion Matrix constructed by Decision Tree Classifier

Decision Tree is used in characterization technique that produces tree structure where every node signifies a test on a trait worth and every branch speaks to a result of the test.Decision tree classifier produces the intermediate results when compared to Naive Bayes , Multinomial Naive Bayes algorithm.

	Support Vector Machine Classifier							
	Unigram		Unigram + POS		Bigram		Bigram + POS	
	Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Positive	1893	107	1767	233	1628	372	1543	457
Negative	205	1885	237	1763	292	1708	382	1618
Accuracy	94.45		88.25		83.42		79.05	

Table 4.3: Confusion Matrix constructed by SVM Classifier

SVM classifier which classifies the classes in two either positive or negative.When a training dataset is Trained by SVM Classifier.The features used here is Unigram,Bigram in which SVM classifier outperforms the rest of classifiers.

Chapter 5

Conclusion and Future Work

Opinion mining is an area in which an enormous volume of data which is generated by person to person communication. Sentiment analysis has different applications such as customer feedback and marketing. With the assistance of opinion mining, organizations can evaluate their market in public and also they learn what are the necessary changes required to make for next product up gradation. And also techniques to make strategies on their item. People can likewise utilize opinion mining for purchasing the product which they haven't used as they will prefer reviews and sentiments of that product before they purchase. In this thesis, we have utilized managed machine learning techniques such as Naive Bayes classifier, multinomial naive Bayes classifier, decision tree, SVM classifier to know the sentiment of that sentence. Unigram and bigram with substantiated results has been found on twitter dataset, 4000 tweets have been used. By using these classification techniques its been concluded that SVM will outperforms other classifiers till date and it is best used classifier.

It is exceptionally hard to recognize objective and subjective data, a large opinionated words likewise happen in target sentences, so it is extremely difficult to handle these challenges. Ordinarily we see peoples posting the surveys in the blogs or forums with tons of spelling mistakes which our word reference can't discover them and bringing about less exactness of sought yield. Along these lines, part of work must be done in this field for recognizing spam online, word sense disambiguation.

Bibliography

[1] Huang, Jin, Jingjing Lu, and Charles X. Ling. "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.

[2] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schtze.Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.

[3] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREC. 2010.

[4] Jiang, Shengyi, et al. "An improved K-nearest-neighbor algorithm for text categorization." Expert Systems with Applications 39.1 (2012): 1503-1509.

[5]] Davidov, Dmitry, Oren Tsur, and Ari Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.

[6] Sharma, Anuj, and Shubhamoy Dey. "A comparative study of feature selection and machine learning techniques for sentiment analysis." Proceedings of the 2012 ACM Research in Applied Computation Symposium. ACM, 2012.

[7] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[8] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal (2014).

[9] B. Liu. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

[10]] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1135, 2008.

[11] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In Proceedings of the 4th Workshop on Social Network Systems, SNS, volume 11, page 8, 2011.

[12] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference, pages 93102. ACM, 2011.

[13] C. Wagner, S. Mitter, C. Korner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In Proceedings of the WWW, volume 12, 2012.

[14] G. Kontaxis, I. Polakis, S. Ioannidis, and E.P. Markatos. Detecting social network profile cloning. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pages 295300. IEEE, 2011.

[15] A. Wang. Detecting spam bots in online social networking sites: a machine learning approach. Data and Applications Security and Privacy XXIV, pages 335342, 2010.

[16] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B.Y. Zhao. Detecting and characterizing social spam campaigns. In Proceedings of the 10th annual conference on Internet measurement, pages 3547. ACM, 2010.

[17] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In Proceedings of the 26th Annual Computer Security Applications Conference, pages 2130. ACM, 2010.

[18] S. Krasser, Y. Tang, J. Gould, D. Alperovitch, and P. Judge. Identifying image spam based on header and file properties using c4. 5 decision trees and support vector machine learning. In Information Assurance and Security Workshop, 2007. IAW07. IEEE SMC, pages 255261. IEEE, 2007.