Sentiment Analysis using Modified Naïve Bayes Classifier

Vidish Raut[1], Manali Shiurkar[2], Sumit Anglekar [3]
Computer Department, Vidyavardhini's College of Engineering & Technology
University of Mumbai, India
vidish.raut@gmail.com
sumitanglekar@rocketmal.com
manali.shiurkar@gmail.com

***Abstract*— Sentiment Analysis (SA) is an ongoing field of research in text mining field. SA is the computational treatment of opinions, sentiments and subjectivity of text. Sentiments are expression of one's words in a sentence. Hence understanding the meaning of text in the sentence is of outmost importance to people of various fields like customer reviews in companies, movie reviews in movies, etc. It may involve huge text data to analyze and it becomes totally unviable for manually understanding the meaning of sentences. Machine Learning Classifier algorithm should be used to classify the sentiment of the text data. We have used supervised machine learning algorithm. By using appropriate training data to train our two different classifiers namely Naïve Bayes, Baseline and Maximum Entropy, we can simplify the task of text classification. In general, we accept a keyword by user as input, fetch tweets related to that keyword from twitter, run classifier on that fetched data and display classified results of twitter tweets in terms of positive, negative and neutral along with the count represented by charts.**

***Keywords*— Text Classification, Machine Learning, Classifier, Sentiment Analysis, Naive Bayes.**

I. INTRODUCTION

Sentiment analysis provide a simple, fast and efficient way to understand large amount of such data and help us to take business related decision quickly. This can provide companies a powerful tool to understand customers and their views and help them to provide better service/products.

Sentimental Analysis is considered to be the future of Ad optimization. Growing availability of opinion rich resources like online review sites, blogs, social networking sites have made this ―decision-making process‖ easier for us. With explosion of Web 2.0 platforms consumers have a soapbox of unprecedented reach and power by which they can share opinions. Major companies have realized these consumer voices affect shaping voices of other consumers. Sentiment Analysis thus finds its use in Consumer Market for Product reviews, marketing for knowing consumer attitudes and trends, Social Media for finding general opinion about recent hot topics in town, Movie to find whether a recently released movie is a hit. The future might see applications wherein a system gauges the human emotion through sensory means and then creates an environment that helps improve the human life in general.[13]

Twitter is a popular micro blogging service where users create status messages (called ―tweets‖). These tweets sometimes express opinions about different topics. There has been a large amount of research in the area of sentiment classification. Traditionally most of it has focused on classifying larger pieces of text, like reviews [1]. Tweets (and micro blogs in general) are different from reviews primarily because of their purpose: while reviews represent summarized thoughts of authors, tweets are more casual and limited to 140 characters of text. Generally, tweets are not as thoughtfully composed as reviews. Yet, they still offer companies an additional avenue to gather feedback.

This experiment's goal is simple: use a Naive Bayes classifier for sentiment analysis, and figure out what we can do to boost its accuracy.

A. DEFINING THE SENTIMENT

Sentiment analysis is a natural Language Processing and Information Extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analyzing a large number of documents. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document.[13]

Table1. Example Tweets [18]

Sentiment	Keyword	Tweet
Positive	Football	Dammmmm I Love Football
Neutral	Airplane	Comes 8 a clock, phone going on airplane mode.
Negative	Pep Guardiola	Pep Guardiola to resign as Barcelona boss

B. CHARACTERISTICS OF TWEETS

Twitter messages have many unique attributes:

1) *Length:* The maximum length of a Twitter message is 140 characters. This is very different from the previous sentiment classification research that

focused on classifying longer bodies of work, such as movie reviews.

2) *Language model:* Twitter users post messages from many different media, including their cell phones. The frequency of misspellings and slang in tweets is much higher than in other domains.
3) *Domain :* Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic. This differs from a large percentage of past research, which focused on specific domains such as movie reviews.[18]

C. APPROACH

The complexity of the problems varies from high to low. So some problems are easily solvable like World Knowledge and some are difficult like Negation. For this purpose various algorithms like Naive Bayes, Baseline and Maximum Entropy at available at our disposal.

Steps for analyzing the sentiments in the sentence:
1. Firstly we need to decide the classifier algorithms and have an appropriate data for training.
2. Preprocess and label the data.
3. Prepare the data for training.
4. Train the classifier with the help of libraries such as NLTK, libsvm etc.
5. Make predictions by giving new test data to the trained classifier. [17]

D. TEXT CATEGORIZATION

Text categorization is the task of assigning a Boolean value to each pair (dj , ci) ∈ D × C, where D is a domain of documents and C = {c1 , . . . , c|C| } is a set of predefined categories. A value of ***T*** assigned to (dj , ci) indicates a decision to file dj under ci ,while a value of ***F*** indicates a decision not to file dj under ci. [2]

In Machine Learning terminology, the classification problem is an activity of supervised learning, since the learning process is ―supervised‖ by the knowledge of the categories and of the training instances that belong to them. [2]

E. MACHINE LEARNING CLASSIFIERS

1) *Naive Bayes:* A Naive Bayes classifier is a well-known and practical probabilistic classifier and has been employed in many applications. It assumes that all attributes (i.e., features) of the examples are independent of each other given the context of the class, i.e., an independence assumption. It has been shown that Naive Bayes under zero-one loss performs surprisingly well in many domains in spite of the independence assumption [5]. In the context of text classification, the probability that a document ***dj*** belongs to a class ***c*** is calculated by the Bayes' theorem as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2) *Maximum Entropy:* The idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint [19]. MaxEnt models are feature-based models. In a two class scenario, it is the same as using logistic regression to find a distribution over the classes. MaxEnt makes no independence assumptions for its features, unlike Naive Bayes. The model is represented by the following:

$$P(c|d,\lambda) = \frac{exp\ \Sigma i \lambda i\ fi(c,d)}{\Sigma\ exp\ \Sigma i \lambda i fi(c',d)}$$

In this formula, *c* is the class, *d* is the tweet, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the *λi's* so as to maximize the conditional probability.

3) *Baseline:*

In this approach, we have used the positive and negative keyword list and for each tweet, we count the number of positive and keywords that appear. This classifier returns the polarity of the highest count. If there is a tie, neutral polarity is returned.

II. MODIFIED NAIVE BAYES ALGORITHM

It turns out that dealing with negations (like "not great") is a pretty important step in sentiment analysis. A negation word can affect the tone of all the words around it, and ignoring negations would be a pretty big oversight. Bigrams, unfortunately, require too much training data, so we've got to find a better way to consider negation terms. Here's how we do it: if we see a negation (like "not", "never", "no", etc), we just add an exclamation point to the beginning of every word after it! The sentence "This movie was not great" turns into "This movie was not !great", and the token "!great" gets stored in our Bayes classifer as having appeared in a negative review. "Great" appears in positive reviews, and its negation, "!great" appears in negative ones. The only modification you need to make is in your tokenizer (the code that goes through the text, pre-processes it, and splits it up in to words or "tokens").

A. LITERATURE REVIEW

Sr. No.	Year	Paper	Description
1.	2002	Thumbs up? Sentiment Classification using Machine Learning Techniques	This paper deals with the problem of classifying documents not by topic, but by overall positive, negative or neutral using Naive Bayes, SVM and Maximum Entropy.
2.	2005	Using Appraisal Group for Sentiment Analysis	It presents a new method for sentiment classification based on extracting and analyzing appraisal groups such as ―very good‖ or ―not terribly funny‖. An appraisal group is represented as a set of attribute values in several task - independent semantic taxonomies, based on Appraisal Theory.
3.	2005	Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis	It presented a new approach to phrase level sentiment analysis that first determines whether an expression is neutral or polar expressions. With this approach, the system is able to automatically identify the contextual polarity for a large subset of sentiment expressions, achieving results that are significantly better than baseline.
4.	2007	Automatic Sentiment Analysis in Online Text	The paper consider the emotions as a classification task: their feelings can be positive, negative or neutral. A sentiment isn't always stated in a clear way in the text; it is often represented in subtle, complex ways. Besides direct expression of the user's feelings towards a certain topic, he or she can use a diverse range of other techniques to express his or her emotions.
5.	2008	Opinion Mining and Sentiment Analysis	This paper covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems.
6.	2010	Twitter as a Corpus for Sentiment Analysis and Opinion Mining	It uses data from micro-blogging site like Twitter and shows how to automatically collect a corpus for sentiment analysis and opinion mining purposes. It perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, it build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document.
7.	2011	Lexicon-Based Methods for Sentiment Analysis	The study presents a lexicon-based approach to extracting sentiment from text.
8.	2013	Unsupervised Sentiment Analysis with Emotional Signals	The authors propose to study the problem of unsupervised sentiment analysis with emotional signals. They incorporate the signals into an unsupervised learning frame work for sentiment analysis. In the experiment, they compare the proposed framework with the state-of-the-art methods on two Twitter datasets and empirically evaluate their proposed framework to gain a deep understanding of the effects of emotional signals.
9.	2014	Comparing and Combining Sentiment Analysis Methods	The study aims at presenting comparisons of popular sentiment analysis methods in terms of Coverage and agreement.

III. RESULT

Result for a particular day:

Fig.1 Using Naïve Bayes

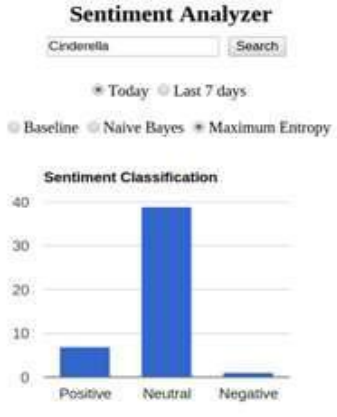

Fig.2 Using Maximum Entropy

IV. CONCLUSION

This paper provides a detail study on various machine learning classifiers. It provides for theoretical evidence that Modified Naïve Bayes algorithm improves the accuracy to 85%.

ACKNOWLEDGMENT

This work was influenced by countless individuals whom we were fortunate enough to meet during the project duration. We are thankful to **Prof. Sangita Chaudhari** for helping us, giving us good ideas for improving our work. Also, we are eager and glad to express our gratitude to the Head of the Computer Department **Dr. Swapna Borde**, for her approval of this project. We are also thankful to her for providing us the needed assistance, detailed suggestion and also encouragement to do the project.

REFERENCES

- Bo Pang and Lillian Lee, Shivakumar Vaithyanathan. ―Thumbs up? Sentiment Classification using Machine Learning Techniques‖. Appears in Proc. 2002 Conf. on Empirical Methods in Natural Language Processing (EMNLP) .[1]
- Fabrizio Sebastiani. ―Machine Learning in Automated Text Categorization‖.[2]
- Vladimir Vapnik(1995) ―Support-Vector Networks. AT&T Bell Labs, Hohndel, NJ 07733, USA.[3]
- Basu, C. Watters, and M. Shepherd(2002).‖ Support Vector Machines for Text Categorization.Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03).‖ [4]
- P. Domingos and M. J. Pazzani, ―On the Optimality of the Simple Bayesian Classifier under Zero-One Loss,‖ Machine Learning,vol. 29, nos. 2/3, pp. 103-130, 1997.[5]
- Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng(2006). ―Some Effective Techniques for Naive Bayes Text Classification‖. (Knowledge and Data Engineering, IEEE Transactions on volume 18, issue 11, 2006) [6]
- Fabrice Colas and Pavel Brazdil. ―Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks.‖ [7]
- Joachims, T. ―Text categorization with support vector machines: Learning with many relevant features.‖ European Conference on Machine Learning (ECML), 1998. [8]
- Susan Dumais. ―Using SVM for text categorization. (Decision Theory and Adaptive Systems Group Microsoft Research)‖ [9]
- S. Rasoul Safavian and David Landgrebe."A survey of Decision Tree methodology". [10]
- Daniela XHEMALI, Christopher J. HINDE and Roger G. STONEIJCSI. International Journal of Computer Science Issues, Vol. 4, No. 1, 2009[11]
- Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.[12]
- Sentiment Analysis: A Literature Survey by Subhabrata Mukherjee-IIT-Bombay [13]
- Kwok, J.T-K. (1998)‖ Automated Text Categorization Using Support Vector Machine.‖ Proceedings of the International Conference on Neural Information Processing (ICONIP).[14]
- Rennie, J.D.M. and R. Rifkin. (2001). ―Improving Multiclass Text Classification with the Support Vector Machine.‖,May 23, 2002 [15]
- http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_ex ports/lguo/decisionTree.html[16]
- https://cloud.google.com/prediction/docs/sentiment_analysis [17]
- Ravikiran Janardhana: ―Twitter Sentiment Analysis and Opinion Mining‖[18]

K. Nigam, J. Lafferty, and A. Mccallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61–67, 1999. [19]