ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

Sentiment Analyzer using Punjabi Language

Anu Sharma

Assistant Professor, Department of Computer Science, DAV College, Chandigarh, India

ABSTRACT: Textual Information in the "Social Media World" can be broadly categorized into two main types: **Facts** and Opinions. Opinions are usually subjective expressions that describe people sentiments, feelings towards entities, events and their properties. Sentiment analysis tracks the mood of the speaker or writer about a particular product or entity. In this paper, an approach is proposed for automatically extracting the movie reviews in Punjabi language from web pages, by using basic NLP technique like N-gram (Unigram, Bigram). The System divides the movie reviews in two categories: Positive and Negative. The System provides an accuracy of 75% on multi-category dataset.

KEYWORDS: Sentiment Analysis, Sentiment Analyzer (SA), Opinion Mining, Natural Language Processing (NLP), Naive- Bayes, N-gram

I. INTRODUCTION

Sentiment Analysis or **Opinion Mining** is a type of Natural Language Processing. It involves building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Automated opinion mining often uses machine learning, a component of Artificial Intelligence. From last few years, web has changed dramatically. People have started expressing their views online via blogs, websites, forums, groups, etc. Online reviews are becoming an important source for different people. If any person wants to buy a product he/she should check the products reviews first. Similarly, if a person wants to watch a movie in theatre, he/she collects information regarding the movie reviews through friends, family members and even through web. Today Web is playing pivotal role in common man's life. Even, nowadays, companies are depending on web for information regarding market conditions, competitors.

This system is performing sentiment analysis at sentence level. Sentiment analysis can be done at word level, sentence level and document level. The main aim of sentence level is to determine whether a sentence is positive or negative. There are several challenges in sentiment analysis. For e.g. a word is considered to be Positive in one situation, may be considered as negative in another situation. Take the word **"long"**. If a customer says that laptop battery backup is long, then it considers as Positive Opinion. But if a customer says hat laptop startup time is long, then it considers as Negative Opinion.

Hatzivassiloglou and McKeown were the first to address the problem of acquiring the prior polarity (semantic orientation) of words. Since then this has become a fairly active line of research in the sentiment community with various techniques being proposed for identifying prior polarity. Turney and Littman use statistical measures of word association.

Paper is organized as follows. Section II describes introduction regarding Punjabi Language. After introduction work of different researchers, of same field has been defined in Section III. After describing work of other researchers, Section IV contains different approaches and actual work of the author has defined in Section V. In this Section two main phases are defined: a) **Training Phase** b) **Testing Phase**. Section VI contains results and conclusion.

II. PUNJABI LANGUAGE (ਪੰਜਾਬੀ)

Punjabi is an Indo-Aryan language spoken by 130 million native speakers worldwide, making it the 9th most widely spoken language in the world. It is the 11th most widely spoken in India. The influence of Punjabi as a cultural language in the Indian Subcontinent is increasing day by day due to Bollywood. Most Bollywood movies now have Punjabi vocabulary mixed in, along with a few songs fully sung in Punjabi [3]. It's strange that very little work has

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

been done in this field. So author has tried to develop an algorithm which performs sentiment analysis for Punjabi Language.

III. LITERATURE SURVEY

Rudy Prabowo and Mike Thelwall [14] They used three approaches viz. **rule based, Support Vector machine and hybrid**. In rule based approach, a rule is an antecedent and its associated consequent is in 'if then' relationship. A consequent represents a sentiment that is either positive or negative. There are various rule based classifiers like GIBC, RBC, SBC and IRBC. GIBC is **General Inquirer Based Classifier** which has 3672 pre classified rules. Out of which 1598 are positive and others are negative. It is applied to classify document. IRBC is the **rule based classifier** in which a second rule set is built by replacing each proper noun found within each sentence with '?' or '#' to form a set of antecedents, and assigning each antecedent a sentiment. SBC is the Statistics Based classifier.

Amitava Das and Sivaji Bandyopadhyay, ICCPOL-10 [3], used the support vector machine (SVM) to developed system for opinion polarity classification on news text in Bengali. They used Bengali SentiWordNet. They have gathered corpus of the experiment from Bengali newspaper sites. They classified news corpus into two types. News reports that aim to objectively present factual information categorized as type 1 whereas opinionated articles in Editorial, forum and letters to the editor categorized as type 2. They developed classifier to mark the sentences which include opinionated words. If any sentence has included opinionated words and theme phrases then they considered sentence as subjective.

To extract the features from the sentence, they used SVM (Support Vector Machine) approach. They used POS tagger to extract the opinion bearing words in sentences. Opinion words in the sentence are mainly adjectives, adverbs, noun and verbs. They have also made list of functional words. Function words are high frequency words and they have no or very less opinionated information.

Faraaz Ahmed, Barath Ashok, Saswati Mujherjee, Meenakshi Sundaran, Murugeshan, Ajay Sampath, ICON-2008[4] They proposed a feature based sentiment classification method. They had used Monty tagger on document of review to extract the part of speech of information. They build the polarity term list. Polarity may be positive or negative. For example "excellent" is positive term where as "bad" is the negative term. Next they extracted feature from review using n-gram and associated them with polarity terms in the review. They found that polarity depended on feature. e.g. The price is too high, which makes it unaffordable. Here the term high has negative polarity. All the polarity term which is context depended called as Local polarity terms. They got a list of polarity terms consisting of positive and negative polarity terms from general inquirer, a publicly available resource. They were extracted adjectives or adverbs from training set and assigned polarity value from global and local polarity list. They also checked the modifiers (such as "yet", "although", "but") effect on the polarity of the sentence. The set of modifiers were categorized as intensifiers and diminishers. For example, if the negation occurred then system changed the polarity to its opposite meaning using WordNet's antonyms. The polarity values of terms in both global and local lists were identified base on the ration terms occurrence in positive or negative reviews against the total number of occurrence. When system has found intensifier with a polarity term then the system has incremented the polarity value term by 3. Similarly for diminisher system has decremented the count of polarity term by 1.

Pang et al.(Pang et al., 2002)[26] used the traditional n-gram approach along with POS information as a feature to perform machine learning for determining the polarity. They used Naive Bayes Classification, Maximum Entropy and Support Vector Machines on a threefold cross validation. Different variations of ngram approach like unigams presence, unigrams with frequency, unigrams+bigrams, bigrams, unigrams + POS, adjectives, most frequent unigrams, unigrams + positions. 82.9% which was obtained in unigrams presence approach on SVM.

Minqing Hu and Bing Lu, AAAI-2004[30] They proposed a method for feature-based opinion summarization of customer reviews of product sold online. They performed this task in two steps. First they identified the features of the product that customer expressed opinion on and then ranked the features according to their frequencies that they appeared in the reviews. Secondly they counted the number of positive and negative reviews or opinion. The input to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

their system was a product name and an entry page for all the reviews of the product. The output was the summary of the reviews.

Thresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce wiebe, yejin choi, Claire cardie, Ellen Riloff, Siddharth Patwardhan, Riloff-2005 [19], developed system capable for supporting natural language processing applications by providing information about the subjectivity in the document. They developed batch and interactive system for opinion finder. In batch mode, system take a list of documents to process where as in interactive mode allows user to query online news sources for the document to process.

The system architecture was one large pipeline which further divided into two parts. The first part was made perform mostly general purpose documents processing like tokenization and part-of-speech tagging. The second part performed the subjectivity analysis.

For the first part the pipeline of system, they used Sundance partial parser to get semantic class tags, identify named entities and match extraction patterns that correspond to subjectivity language. For tokenize, sentence split and part of speech tag they used OpenNLP[1] 1.1.0. The second part of their subjectivity analysis system has four components: 1) Subjective sentence classification, 2) Speech events and direct subjective expression classification, 3) Opinion source identification and 4) Sentiment expression classification. They used naïve bayes classifier for subjective sentence classification. They trained classifier by subjective and objective sentences. Second component of subjectivity analysis identified speech events like said, according to etc. and direct express like fear, is happy etc. For the third component of subjectivity analysis they combined Conditional random field sequence tagging model and extraction pattern learning to get the source of speech events and direct subjective expression. The third component was trained using MPAQ Opinion corpus.

They have developed two classifiers using BoosTexter for sentiment expression classification. The first classifier focused on identifying sentiment expression and second classifier took the sentiment expressions and identifies positive and negative.

Amitava Das and Sivaji Bandyopadhyay, IEEE-09[1] They developed the subjectivity detection system which was evaluated on Multi Perspective Question Answering (MPQA) corpus as well as on Bengali corpus. They defined Opinion as private state. Subjective remarks come in various forms including opinions, rants, allegations, accusations, suspicions, humor and speculation. They developed the theme subjectivity detection system based on rule base technique. This worked in two stages:

(a) First captured discourse level opinion theme in terms of thematic expressions.
(b) Then examined the presence of thematic expression as an opinion constituent (Subject- Aspect evaluation).

IV. APPROACHES

Naïve Bayes Classifier: Naive Bayes classifier is simple but effective learning system. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, Naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently for supervised learning. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the Naive Bayes model without believing in Bayesian probability or using any Bayesian methods.

Each piece of data that is to be classified consists of a set of attributes, each of which can take number of possible values. The data are then classified into a single classification. It is based on Probabilistic reasoning.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

To identify the best classification for a particular instance of data (*d*1, . . .,dn), the posterior probability of each possible classification is calculated:

P(ci| d1, . . ., dn)	Eq.(1)

where *ci* is the *i*th classification, from a set of |*c*| classifications. The classification whose posterior probability is highest is chosen as the correct classification for this set of data. The hypothesis that has the highest posterior probability is often known as the maximum a posteriori, or MAP hypothesis. In this case, we are looking for the MAP classification.

To calculate the posterior probability, we can use Bayes' theorem and rewrite it as

$\frac{P(d_1,\ldots,d_n\|c_i)\cdot P(c_i)}{P(d_1,\ldots,d_n)}$	Eq.(2)

Because we are simply trying to find the highest probability, and because *P*(*d*1, . . ., *dn*) is a constant independent of *ci*, we can eliminate it and simply aim to find the classification *ci*, for which the following is maximized:

$P(d_1,\ldots,d_n\|c_i)\cdot P(c_i)$	Eq.(3)

The naïve Bayes classifier now assumes that each of the attributes in the data item is independent of the others, in which case *P*(*d*1, . . ., *dn*|*ci*) can be rewritten and the following value obtained:

$P(c_i)\cdot\prod_{j=1}^{n}P(d_j\|c_i)$	Eq.(4)

The naïve Bayes classifier selects a classification for a data set by finding the classification *ci* for which the above calculation is a maximum.

V. ALGORITHM

This algorithm performs Polarity based Classification on data set. Polarity is divided into two parts: **Positive Polarity** and **Negative Polarity**. The system follows two main phase: Training Phase and Testing Phase. The system performs N-gram techniques (Unigram, Bigram and Combination of Unigram and Bigram Technique). Data is collected from different websites. As for Punjabi language no such resource is present. The author collects data from different Punjabi newspapers, blogs. The collected data is called corpus here. Author uses Naive Bayes classifier. In Training phase, the system analyse the paragraph. Then the collected data is differentiated on the basis of movies rating. The movies having 2.5 or more than 2.5 ratings are considered as Positive Polarity data and the reviews having less than 2.5 ratings are considered as Negative Polarity data. After this testing phase will start its work. In starting all the words present in corpus, author assigned the positive and negative frequency to zero by using following equations:

$PF_i = \sum_{Wordi \in Positive\_Corpus} PF_i + 1$	Eq.(5)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

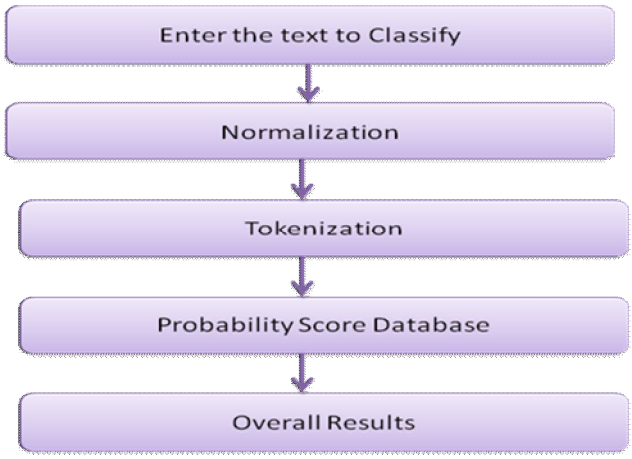
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

$NF_i = \sum_{Wordi \in Negative\_Corpus} NF_i + 1$	Eq.(6)

After this phase second phase testing phase starts. Testing phase contains following steps:



The system calculates the probability using following formula:

$POL\_POS = \sum_{i=0}^{n} POS\_SCORE_i$	Eq.(7)

$POL\_NEG = \sum_{i=0}^{n} NEG\_SCORE_i$	Eq.(8)

At last, if the POL_POS >POL_NEG then the review has Positive Polarity and if the POL_POS < POL_NEG then the review has Negative Polarity.

VI. RESULTS AND CONCLUSION

Author has tried to explore and analyze the Naive Bayes classification methods (N-Gram approach) based on supervised learning. Author have also tried to provide a general framework in order to deal with sentiment analysis (opinion mining) efficiently. We have conducted our experiment on Movie news rated by user.

Table 1: Result Table

Technique	Accuracy
Unigram only	**68%**
Bigram only	**70%**
Unigram + Bigram	**75%**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2014

Finally author concludes that Bigram approach gives better result than Unigram approach and combination of **Unigram + Bigram Technique** provides better accuracy.

REFERENCES

1. Amitava Das, Sivaji Bandyopadhyay (2009). "Theme Detection an Exploration of Opinion Subjectivity",pp 1-6,IEEE-09".
2. Amitava Das and Sivaji Bandyopadhyay (2010)."SentiWordNet for Indian Languages",pp 56-63, AFNLP-10.
3. Amitava Das, Sivaji Bandyopadhyay (2010). "Opinion-Polarity Identification in Bengali", pp- 169-182, ICCPOL-10.
4. Faraaz Ahmed, Barath Ashok, Saswati Mukherjee, Meenakshi Sundaram, Murugeshan, Ajay Sampath (2008)."Effect of Modifiers for Sentiment Classification of Reviews", ICON-08.
5. Lun-Wei Ku, Yu Thing and Liang Hsin-His Chen. (2006). "Opinion Extraction, Summarization and Tracking",pp 100-107, AAAI-06.
6. Minqing Hu and Bing Lu (2004), "Mining and Summarizing Customer Reviews". ACM New York, pages -168-177, KDD'04
7. Minqing Hu and Bing Lu, (2004) "Mining opinion Features in Customer Reviews", pp-755–760, AAAI-04.
8. MPQA Opinion Corpus, http://nrrc.mitre.org/NRRC/publication.html
9. Naïve Bayes Classifier, http://en.wikipedia.org/wiki/Naive_Bayes_classifier
10. N-gram, "http://en.wikipedia.org/wiki/N-gram"
11. NTCIR corpus, http://research.nii.ac.jp/ntcir/index-ex.html
12. Tamara Martin –Wanton ,Aurora Pons-Porrata and Andres Montoyo-Guijarro,Alexandra Balahur (2010) "Opinion Polarity Detection ",pp 483-486,ICAART-10
13. Precision and Recall, http://en.wikipedia.org/wiki/Precision_and_recall
14. Rudy Prabowo and Mike Thelwall (2009)"Sentiment Analysis: A Combined Approach", pp 143-157, ScienceDirect-09.
15. SentiWordNet, http://SentiWordNet.isti.cnr.it/
16. Please see www.jagbani.com
17. Please see www.punjabitribuneonline.com
18. Si Li, Hao Zhang, Weiran Xu, Guang Chen and Jun Guo. "Exploring Combined Multi-level model for Document Sentiment Analysis", pp 4141-4144,IEEE-2010
19. Theresa Wilson, Paul Hoffmann, Swapna Somasundaran ,Jason kessler, Janyce Wiebe Yejin Choi, Claire Cardie ,Elle Riloff <Siddharth Patwardhan(2005)."Opinion Finder:A system for Subjectivity analysis", HLT-Demo , Proceedings of HLT/EMNLP on Interactive Demonstrations, Association for Computational Linguistics,pp 347-354,2005.
20. WordNet, http://wordnet.princeton.edu/
21. Xiaowen Ding, Bing Liu, Philip S. Yu, (2008)"A holistic Lexicon-Based Approach to Opinion Mining", pp 231-240, WSDM-08.
22. Akshat Bakliwal, Ankit Patil, Piyush Arora,Vasudeva Varma,"Towards Enhanced Opinion Classification using NLP Techniques", Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP),Chiang Mai, Thailand, November 13, 2011 ,pp101-107, IJCNLP 2011.
23. Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pp 1,Chiang Mai, Thailand, November 13, 2011.
24. Vishal Goyal, Ankur Rana, Vimal K. Soni,(2011) "Renaissance of Opinion Mining",pp 60-67,(2011).
25. Kushal Dave, Steve Lawrence, and David M. Pennock.Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Pp 519–528,2003.
26. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 79–86, 2002.
27. Bo Pang and Lillian Lee. A sentimental education:Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the ACL,
28. pp 271–278, 2004.
29. Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. pp 417–424,2002.
30. Vasileios Hatzivassiloglou and Kathleen Mckeown (1997)"predicting the semantic orientation of adjectives" pp 174-181, EACL '97.
31. Minqing Hu and Bing Liu, (2005) "opinion Extraction and summarization on web", pp 1621-1624 , AAAI-05.

BIOGRAPHY

Anu Sharma is an Assistant Professor in the Department of Computer Science, DAV College, Chandigarh, India. She received her M.phil(CS) degree in 2013 from DCS, Punjabi University, Patiala, Punjab, India. She received her Master of Computer Application (MCA) degree in 2011 from DCS, Punjabi University, Patiala, Punjab, India. Her research interests are Natural Language Processing(Sentiment Analysis, Information Retrieval, etc)