Opinion Mining Using Polarity Classification Algorithm

K.Sandhya[#1], V.Lavanya Rani[#2], P.Padma Kumari[#3,] U.Mounika[#4]
[#1, 2,, 3, 4] *Students CSE Department, Lendi institute of engineering and technology, Vizianagaram, Andhra Pradesh, India.*
[1]sandhyaksukanya@gmail.com, [2]lavanyavadapalli135@gmail.com, [3]padma.friendlyto@gmail.com, [4]mounikauppalapati22@gmail.com.

***Abstract*—Sentiment analysis is a broad research area in academic as well as business field. The term sentiment refers to the feelings or opinion of the person towards some particular domain. Hence it is also known as opinion mining. It leads to the subjective impressions towards the domain, not facts. It can be expressed in terms of polarity, reviews or previously by thumbs up and down to denote positive and negative sentiments respectively. Sentiments can be analyzed using Natural Language Processing(NLP), statistics or machine learning techniques. The primary issues in previous techniques are classification accuracy, data sparsity and sarcasm, as they incorrectly classify most of the feeds with a very high percentage of feeds incorrectly classified as neutral. To overcome this we propose an algorithm called Polarity Classification Algorithm for feeds classification based on a hybrid approach. The proposed method includes various pre-processing steps before feeding the text to the classifier. Experimental results show that the proposed technique overcomes the previous limitations and achieves higher accuracy when compared to similar techniques.**

***Keywords*—Sentiment analysis, Opinion Mining, Polarity Classification, NLP.**

I. INTRODUCTION

Opinion Mining is one of the most powerful areas of data mining- helps organizations in discovering valuable insights into the needs and preferences of their customers. To be specific, in a given piece of text, opinion mining aims to identify the part which is expressing the opinion and what is being communicated. A more rigorous and detailed analysis of opinions in the public market place delineates aspects such as polarity (negative, positive, or neutral), subjectivity of the opinions proffered, polarity strength of the contextual piece of text. There are two main approaches for analyzing sentiments, namely Machine Learning Approach and Lexicon-based Approach. Machine Learning approach applies machine learning algorithms with linguistic features and can be implemented using either supervised learning or unsupervised learning methods. On the other hand, Lexicon-based approach depends on a sentiment lexicon, i.e. an assortment of known and precompiled sentiment terms. Lexicon-based approach can be implemented using two methods, i.e. dictionary-based approach and corpus-based approach. Additionally, Hybrid Approach, i.e. a combination of both Machine Learning and Lexicon-based approach can also be used to discover the true meaning and emotion behind many of the reviews and comments posted by customers online. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer's preferences, which could prove valuable for economic or marketing research.

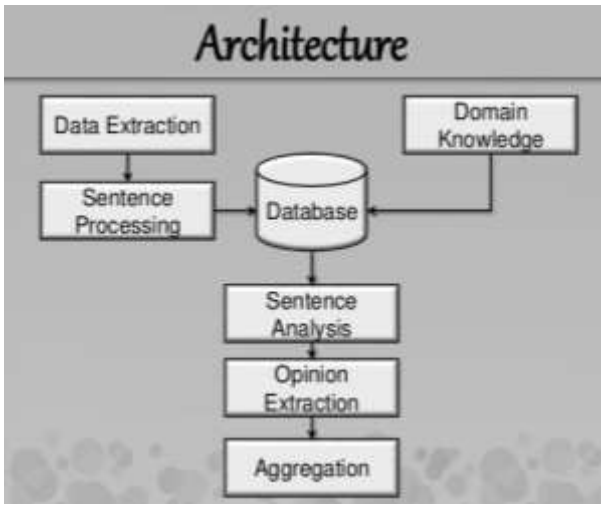

Fig-1:Opinion mining Architecture

Data classification is the process of organizing data into categories for its most effective and efficient use. Classification models predict categorical class labels. The data classification Process includes two steps:
Building the classifier model
Using classifier for classification

Classification routines in data mining also uses a variety of algorithms—and the particular algorithm used can affect the way records are classified. Opinion Mining or Sentiment Analysis is a Natural Language Processing and Information Extraction task that identifies the user's views or opinions explained in the form of positive, negative or neutral comments and quotes underlying the text. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining.

II. EXISTING SYSTEM

There are some best classification algorithms for mining the opinions.Among them Naïve Bayesian (NB) and Support Vector Machine (SVM).The results show that SVM has maximum accuracy (93.25 %) . Four machine learning algorithms are compared: Naïve Bayes, SVM, VFI (Voting

Feature Intervals) and Hyperpipes. The best precision achieved was 79% using Hyperpipes with non-stemmed, non-rooted, mutually deducted feature vectors containing 2000 features. The Result shows that SVM, with almost 90% accuracy, is a more accurate classifier compared to NB in categorization tasks.

Naive Bayes, maximum entropy and Support Vector Machine algorithms and shown the good results as comparable to other ranging from 71 to 85% depending on the method and test data sets. SVM turns out to be the most stable learning algorithm, and also provided the best accuracy (62 %).

DEMERITS:

Limitations occurred when clustering is applied to high dimensional data are

1. Better accuracy but precision and recall will keep very low.
2. Data Sparsity Issue.
3. Incorrect Classification of feedbacks.

To tackle these problems,Unsupervised learning algorithms are used for data classification.

III. PROPOSED SYSTEM

Customers' opinions and comments voiced over social media through blogs, social networking websites, RSS feeds, and discussion forums are not just limited to their friends and family, but can be accessed by anyone around the world. This huge volume of opinionated data created is very useful for individuals as well as for organizations because of the power of influence and the social conditioning of influence. A lot of what we do as daily consumers, our perceptions, our beliefs and the varied choices that we make in the marketplace are largely conditioned by others and our networks and community influencers. We often seek out the opinions of others before making purchase- and especially big ticket ones-decisions. Individuals often seek advice or opinions of relatives, family members, office colleagues or friends whereas organizations and businesses tend to use focus groups, market research surveys, consultant advisors and opinion polls. The proposed system applies a variant of techniques for feed analysis and classification. This involves pre-processing steps and a hybrid scheme of classification algorithms. Pre-processing steps include: removal of URLs, hash-tags, username & special characters; spelling correction using a dictionary; substitution of abbreviations & slangs with expansions, lemmatization and stop words removal. The proposed classification algorithm incorporates a hybrid scheme using an enhanced form of emoticon analysis, SentiWordNet analysis and an improved polarity classifier using a list of positive/negative words.

The main goal of our project is to improve the accuracy of text classification and resolve the data sparsity issues. The core idea is to pre-process the raw data and perform different transformations to remove the slangs, grammatical mistakes, abbreviations and other noise and then feed it to the classifier. The feeds are used as input items. The proposed system is basically composed of three main modules. The first module is data acquisition, a process of obtaining feeds from web; the second module performs pre-processing and transforms the feeds containing real valued features or arbitrary components and refines them into a stream pattern that can be easily used for subsequent analysis. The last component applies different classification techniques in a pipelined way which classifies the feedbacks into positive, negative or neutral.

MERITS:

- Tests the accuracy of sentiment identification on feedback datasets, and produces an average harmonic mean of 83.3% and accuracy of 85.7%, with 85.3% precision and 82.2% recall
- Resolves the data sparsity issue using domain independent techniques.
- Comparison with other techniques to prove the effectiveness of the proposed hybrid approach.

IV. CONCLUSION

It is found from above work that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefit from each other's merits, and finally enhance the sentiment classification performance. The overall accuracy of the polarity classification algorithm is about 73%. Helpful negative reviews can be found with 82% precision and 77% recall. Helpful positive reviews can be found with 74% precision and 64% recall. Unhelpful reviews can be filtered out automatically from the consumer reviews with a high recall rate of about 87% with 73% precision.

REFERENCES

[1]. B.Seerat and F.Azam "Opinion Mining: Issues and Challenges (A survey)"International Journal of Computer Applications (0975 – 8887) Volume 49–No.9, July 2012.

[2]. Balahur, Sentiment Analysis in Social Media Texts, 2013, pp. 120–128, Atlanta,Georgia.

[3]. Ding, X., Liu, B., YuA, P.:HolisticLexicon-Based Approach to Opinion Mining. Proc. of the Int. Conf. on Web Search and Web Data Mining WSDM'2008, New York, NY, USA, 231-240 (2008)
[4]. Liu B. Sentiment analysis and opinion mining. Synth LectHuman Lang Technol 2012
[5]. Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management,pp. 617-624, Bremen, Germany, 2005
[6]. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up Sentiment Classification using Machine Learning Techniques In: Proceedings of EMNLP;2002.p. 79-86.
[7]. Monalisa Ghosh, Animesh Kar, Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using SentiWordNet 3.0, International Journal of Engineering Research and Technology 2013;2- 9
[8]. Park, A. and Paroubek, P. 2010. Twitter as aCorpus for Sentiment Analysis and OpinionMining. LREC 2010

