International Journal of Advanced Computational Engineering and Networking, ISSN(p): 2320-2106, ISSN(e): 2321-2063
Volume-5, Issue-12, Dec.-2017, http://iraj.in

ASPECT LEVEL APPROACH IN OPINION MINING WITH EXTENDED FEATURES

[1]M.MANGAIYARKARASI, [2]N.PUVIARASAN

[1]Research Scholar, Department of Computer and Information Sciences, Annamalai University, Chidambaram
[2]Professor, Department of Computer and Information Sciences, Annamalai University, Chidambaram
E-mail:[1]mjmangaiyarkarsi@gmail.com, [2]npuvi2410@yahoo.in

Abstract - Opinion mining is a way to retrieve opinions through search engines, Web blogs and social networks. It's today's trend that people tend to seek opinion on everything for which we need an efficient Opinion Mining System. In top of our research, we have successfully developed a high-precision opinion mining system which will evaluate opinions from various sources that helps to every level of user. The most successful supervised methods like decision tree have been deployed in our system. With current systems we have learned that location is not considered as a factor in evaluating opinions that will not be helpful in all the cases. Considering this we have taken critics location as an additional factor in evaluating opinions.

Keywords - Opinion Mining; Sentiment Analysis; Information Retrieval; Text Mining; Web Mining

I. INTRODUCTION

The opinion of people has always been an essential information most of us during the decision-making process. Unlike older days, internet and Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics - that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the Internet. Opinion mining in particular for market research requires large quantity of data. With the support of technology available now it is ease to collect required data through social media and other online review systems. It is also necessary to spot the characteristics of the product in addition to the user reviews. There are mainly four predominating problems viz. subjectivity classification, word sentiment classification, document sentiment classification and opinion extraction [4]. As we mentioned earlier the sensible classification of product characteristics as well as reviews will enhance the decision supporting skills of any opinion mining system.

Various techniques exist that can be used for sentiment analysis task [6]. The main approaches are Supervised Methods or Machine Learning and Unsupervised Methods or lexicon-based. Machine learning approach uses dataset for training classifier which will be further applied for defining sentiment of a particular text. The lexicon-based method uses the semantic orientation (SO) of words or phrases to define whether a text is positive or negative.

II. LITERATURE REVIEW

Feature extraction is the primary task in opinion mining i.e., Target subject/object's feature reviewed by author needs to be extracted. The features can be divided into explicit and implicit [10]. Explicit features are direct comments and Implicit were indirect comments. The review may contain infrequent features and that needs Pruning approaches to improve the precision of feature extraction [12]. Same author proposed an idea to extract implicit features by manually labeling the words related to the feature [9]. Lexicon based approaches were another popular approach in in Opinion mining. The idea is to build the lexicon dictionary with known sentimental words. The adjectives and verbs of the sentences will be reviewed carefully to build the lexicon base [11]. With growing technology Machine learning approaches comes into play and document level opinion identification is made possible with this [14]. Turney [16] proposed a successful idea to perform unsupervised classification method that will perform opinion mining with Web Search Engines. A more sophisticated approach developed on mid 90's, which used a WordNet distance based method to determine the opinion orientation of a given adjective [13].

There is another approach known as random walk model over a word relatedness graph to produce a sentiment estimate for a given word. This method uses WordNet synonyms and SentiNet dictionaries to build a word relatedness Graph [5]. The contextual subjectivity is another important consideration in corpus based approach that means that although a word or phrase in a lexicon is marked positive or negative, but in the context of the sentence expression it may have no opinion or have the negative opinion [7].

III. METHODOLOGY

1. Decision Tree

Decision tree is basically the hierarchical decomposition of the data space. The hierarchical decomposition will be based on the attribute value of the node [3]. The hierarchical decomposition will

happen until meeting the classification conditions. The identified leaf nodes will be used for classification. For a given test instance, we apply the sequence of predicates at the nodes, in order to traverse a path of the tree in top-down fashion and determine the relevant leaf node. The portion of the data which is held out is used in order to determine whether or not the constructed leaf node should be pruned or not. Decision tree supports many other ways of classifications. It may not be necessary to identify individual attributes we may need to check the collective attribute value which may be an opinion classification. Such correlated sets of terms and attributes can be identified and decomposed into leaf nodes of decision tree. In Decision Trees the predicate the predicate or the attribute value of the subject is used to decompose the data space hierarchically. The classification rules for decision tree will be calculated from the available training data.

2. Naive Bayes

Naive Bayes is a simple and efficient data mining approach that can be used for opinion classification. Naïve Bayes classifier is a probabilistic approach integrating the Bayes' algorithm that allows to compute probability of features [1] belonging to a label:

P (label|features) = P (label) * P (features|label) / P (features)

Where

P (label|features) is the posterior probability of features belonging to a label (positive or negative)

P (label) is the prior probability of a given label,

P (features|label) is the conditional probability that the particular feature in features appears given label

P (features) is the prior probability of the feature in features. If the features were independent to each other, the Naives can be calculated as below,

$$P\ (features|label)\ approx. = P\ (f1|label) * P\ (f2|label) * \dots * P\ (fn|label) = {}^{n}\pi_{i-1}P(fi|label)$$

$$P\ (label|features) = (label) * {}^{n}\pi_{i=1}(fi|label) / P\ (features)$$

The individual feature is referred as fi(π).

Although Naive Bayes model assumes that features are generated independently of their positions, it still gives good result in real tasks.

The main goal of the classification is to define the label the feature belongs to. Therefore, we do not interested in finding the probability itself, however, the most probable label has to be defined. Naïve Bayes classifier uses the maximum a posteriori (MAP) estimation to define the most probable label $label_{map}$

$$label_{map} - \max_{label\in}[P(label) * {}^{n}\Pi_{i=1}P(fi|label) / P(features)]$$

IV. DATA AND PREPROCESSING

We have deployed supervised methods in our experiment. This requires training dataset. We have collected two datasets for training classifiers. The first dataset is a dataset we introduce here is the reviews about the Samsung Galaxy Note8 mobile. Although n number of reviews available we have taken 280 of them and they were equally positive and negative. We have taken 30 positive reviews and 12 negative reviews from the training data and consider them as testing data. We have taken calculated percentage of data from training data for testing. As so the training data can be evaluated with that percentage of testing data. We have considered the Samsung Note 8 data as it is trendy and good enough to contain variety of opinions. The second dataset were the reviews about the library of Annamalai University. We have taken this subject as to showcase the variety and to establish that our researches can evaluate opinions of any particular subject, not only products. The training data includes total of 140 reviews, which are labeled as positive and negative and 14 among them were tested. We have represented the discussed statistics in the below table **Table 1**.

Data Set	Type	Positive Opinions	Negative Opinions	Total Reviews
Samsung Note 8	Training	140	140	280
	Test	14	7	21
AU Library	Training	110	30	140
	Test	11	3	14

Table 1: Training & Testing Data Statistics

The primary step to be considered on preprocessing is to identify and skip the unnecessary words. Preprocessing is crucial in performance tuning and the computation time. The irrelevant data can slower the learning process and decrease the efficiency of the system in general. We have identified and highlighted the general follow-ups [2] for Preprocessing as specified below:

- Removal duplicates – We may receive similar reviews and that has to be removed to speed up the mining.
- Removal of special characters – The special characters have no impact on opinion mining and it will be additional burden to our experiment. So it has to be identified and kept out of the

evaluation.

- Removal of hyperlinks – Hyperlinks and data reference to that will highly influence the performance. That has to be eradicated to make the classification simple.

Truncating elongated words – Long words like "soooooper", "verrrry" needs to be identified and kept out of the loop to avoid unnecessary classification.

V. RESULTS AND ANALYSIS

We performed our experiments with two different datasets. The first dataset contains the reviews of Samsung Note 8 mobile phone and second dataset contains reviews about library of Annamalai University. Our Opinion Mining System evaluates the reviews using two different approaches Decision Tree and Naïve Bayes. Both datasets were tested with both the approaches. We have developed our own tool to perform this mining and our tool will run successfully with any basic configuration. In Decision tree we have used Boolean logic to generate the predicates and identify the essence of review. With Naïve Bayes approach we will identify the opinions and classify it based on its sentiment to frame a training set. We will identify and frame the training set into our database. With the help of that DB we will evaluate the reviews.

Opinions can be classified differently based on its features and lexicons. We have proposed the most general category as to classify the reviews or opinions to make it simple. Any of our input opinion will fall among these categories,

- Positive – If the majority of opinions appear positive.
- Negative – If the majority of opinions appear negative.
- Neutral – If the opinions equally dispersed with positive and negative features.

However the generated report alone will not ensure the precision of our system. So that we have cross validated it to improve and display the effectiveness of our system. The Classification algorithms were highly effective based on evaluation and estimation of few metrics [15] Listed in Table 2. Moreover the resources adapted by the algorithms and cost impact of that with respect to performance are an important factor that has to be considered.

Performance Metrics
Precision
Feature Score
Recall
Accuracy

As we discussed these metrics decides the success of classification algorithm, let us discuss the way to estimate them.

Accuracy = $\sum(x+y) / \sum(x+y+\tilde{x}+\tilde{y})$

Where,

X denotes the positive opinions which are really identified as positive.

Y denotes the negative opinions which are really negative.

$\tilde{x}$ denotes the negative opinions falsely identified as positive.

$\tilde{y}$ denotes the positive opinions falsely identified as negative.

With similar parameters precision can be estimated as.,

Precision = $x / (x+ \tilde{x})$

With the above precision estimation we can predict the hit rate of positive classifiers. Higher the hit rate, the precision is higher.

Recall = $x/ (x+\tilde{y})$

The recall reveals the classifiers ability to predict the percentage of positive answers out of the expected positive opinions.

$$F1 = 2 * precision * recall / precision + recall$$

The balance between the above two metrics can be estimated using harmonic mean formula given above. By estimating our system with the above metrics we have noted that with small dataset the metrics value goes lower. Higher the dataset the accuracy is more. The highest Mean value we have reached on estimating F1 was 85.20 % with above 5000 words. However with most appropriate words the rate goes higher with smaller data.

CONCLUSION

This paper analyzes two classification algorithms that can be used for opinion mining. With reference to prior researches we have found that majority of opinion mining were more successful with supervised machine learning approaches. We have successfully analyzed both algorithms, tested well and their performances were estimated. Additional to most of the opinion mining researches we have preferred location of the reviewer. We have learned that the opinion varies based from location to location. This will be more helpful to institutions, organizations and researchers in decision making and to improve themselves.

REFERENCES

[1] Vrinda, Dr. Komal Kumar Bhatia, "Opinion Mining using Naive Bayes Classifier",International Journal of Engineering Research & Technology, 2017.

[2] Jingcheng Du, Jun Xu, Hsingyi Song, Xiangyu Liu and Cui Tao,"Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets", Journal of BioMedical Semantics, 2017.

[3] Nur Atiqah Sia Abdullah , Nurul Iman Shaari and Abd Rasid Abd Rahman,"Review on Sentiment Analysis Approaches for Social Media Data", Journal of Engineering and Applied Sciences,2017.

[4] Shailesh Kumar Yadav,"Sentiment Analysis and Classification: A Survey", International Journal of Advance Research in Computer Science and Management Studies, 2015.

[5] Ahmed Hassan, Amjad Abu-Jbara, Wanchen Lu,Dragomir Radev,"A Random Walk–Based Model for Identifying Semantic Orientation", Conference on Computational Linguistics,2014.
[6] Maqbool Al-Maimani, Naomie Salim, Ahmed M. Al-naamany,"Opinion mining: approaches, resources and challenges", Journal of Theoretical and Applied Information Technology, 2014.
[7] Lizhen Qu, Cigdem Toprak, Niklas Jakob, Iryna Gurevych,"Sentence Level Subjectivity and Sentiment ", Analysis Experiments in NTCIR-7 MOAT Challenge, 2008.
[8] A. Ghose, P. Ipeirotis, and A. Sundararajan,"Opinion mining using econometrics: a case study on reputation systems", Annual Meeting-association for computational linguistics, 2007.
[9] B. Liu, M. Hu, and J. Cheng,"Opinion observer: analyzing and comparing opinions on the web", In WWW '05: Proceedings of the 14th international conference on World Wide Web, 2005.
[10] M. Hu and b. Liu,"Mining and summarizing customer reviews", Proceedings of the Tenth ACM SIGKDD International conference on knowledge discovery and data Mining, 2004.
[11] S.-M. Kim and E. Hovy,"Determining the sentiment of opinions", In Proceedings COLING-04, the Conference on Computational Linguistics, Geneva, 2004.
[12] M. Hu and B. Liu,"Mining opinion features in customer reviews", In Proceedings of AAAI-04, the 19th National Conference on Artificial Intelligence, San Jose,2004.
[13] Jaap Kamps, Maarten Marx, Robert J. Mokken, Maarten de Rijke,"Using WordNet to Measure Semantic Orientations of Adjectives", In Proceedings of the International Conference on Language Resources and Evaluation (LREC),2004.
[14] B. Pang, L. Lee, and S. Vaithyanathan,"Thumbs up? Sentiment classification using machine learning techniques", In EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Morristown, NJ, 2002.
[15] Sergio A. Alvarez,"An exact analytical relation among recall, precision, and classification accuracy in information retrieval", 2002

★★★

.