International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 2 238 - 242

Mining High Dimensional Web Content With Sentimental Analysis: A Proactive Approach

Mrs. S.Yamini,
Research Scholar,
Department Of IT,
Bharath University, Chennai.
yaminianitha@yahoo.co.in

Dr. V. Khanna,
Research Supervisor,
Dean, Centre for Information,
Bharath University, Chennai.
drvkannan62@yahoo.com

Dr. Krishna Mohanta,
Professor, Department Of CSE,
Sri Ramanujar Engineering College,
Chennai.
krishnamohanta@gmail.com

***Abstract*--** Sentiment analysis or opinion mining has a humongous scope in the field of digital marketing. Many research ideas have evolved in this field of engineering over the past decades. The major task of proposing sentiment analysis in mining is to systematize the detection of opinions, attitudes and the feelings expressed. These approaches however have some setback in certain scenarios. Rather than directly expressing the feelings sometimes a person might use diverse strategies to express emotions which might be positive, negative or neutral. One word which was viewed positive in a scenario might be regarded as negative in another situation. Such circumstances would interrogate the reliability of sentimental analysis. Our researches aim at alleviating the challenges in sentimental analysis and deliver a tool that is effective and reliable.

Keywords *– Sentiment analysis, Text categorization, opinion mining, text analysis etc…*

*****

I. INTRODUCTION

Online contents have seen a humongous growth in dimensionality over the past decades. These are not alarmingly increasing in terms of size but also in the variety of services they offer. Managing these bulk data has been always a challenging aspect with our current technologies available for the same. Day by day new users are integrated and existing users indulge in various activities which generates more and more new data thereby increasing the total size to a colossal amount. These new contents are mostly blog entries, opinions about products or some services. Mining these data would be an influential part in data mining as these user feedback data would impact other user's choice which would have an important cradle of info for any industry to take into consideration while developing marketing enhancement strategies.

The analysis of sentimental data found online would be vibrant for any organization concerned about customer fulfilment and quality governance. Attaining user response means probing them with surveys on various features the organization would be interested in.however there are a few hardships in this approach too. Making a survey, defining an appropriate format for the same, circulation, timing of release, willingness of the people who take part etc… would take a vital role on the same. One of the best ways to extract opinions accurately is through analyzing blogs; a platform where users express remarks about a topic or else would convey their private thoughts or otherwise would request other users to express their opinions about the same.Integrationof such methods in current search engines would empower users to specificallyanalyze the documents encompassing data that are for or else in contradiction of a topic.

It might resemble alike to another field of study which is being vigorously researched and is known as topical categorization of data. However these two are dissimilar to each other in quite a few aspects. The difficulties in sentiment or opinion analysis is echoed by the letdown of all the earlier attempts to conquer exactness alike those that were previously achieved in topical categorization [1]. This largely ascends owing to the point that in sentiment analysis the overall opinion expressed might be dissimilar from the opinion expressed in individual sentences. This sarcasm can be commonly seen only in film reviews. Contemplate a simple technology based on sparse vector of occurrence counts of words [2]. Probability of it performing wretchedly is more while reviewing a good gory, horror movie, since such a review will be complete with words having contrary opinions in the parts where it reveals about the design of the screenplay. This statement is also sustained byTurney's(2002) [3] work on classification of reviews.

II. PREVIOUS WORK

An earliest effort in this field of study was to classify the category of texts, for illustrating a particular category (Karlgenand Cutting, 1994; Finn et al., 2002). The preliminary tactics in opinion mining mostly utilized philological heuristics, unambiguouslist of pre-nominated words and other similar technologies that necessitate use of professionals knowledge and may not harvest the most likely fallouts in all cases as described in Bo Pang et al.,2002[4].

An initial task to systematize the opinion mining was perceived in the work of Turney(2002)[3]. Tumney utilized the shared info between a phrase and the words "excellent" and "poor" as a measurementfor categorization. This info was obtained on the basis of numbersassembledby a search engine. However, the actualadvancement in this field came with the work of Bo Pang et al.(2002)[4]. Enchanting the triumph of the managed learning techniques in the province of text classification as an encouragement, they practiced it in movie reviews and obtained anabundantdevelopmentbetter than the earliertactics.

The word sentiment used in orientation to the systematic analysis oftext and pursuing of the exploitive conclusionsthat appears Das and Chen et al, [6].Consequently, this model was embraced and improved byTurney [3] and Pang et al. [1]. This idea was conceded on by Nasukawa & Yi [5] and Yi et al. [7]. These proceedings combined might illuminate the reputation of pinion mining among groups self-recognizedas engrossed on Natural Language Processing.A considerable number of articles that were stating about opinion mining focused on aprecise application of categorizingcustomer review as to their divergence either positive or negative.

III. FUNCTIONS OF OPINION MINING

Opinion mining can be practically categorized into three major tasks like development of linguisticassets, sentimentalanalysis, and opinion summarization.Martin J.R in [8] depicts the appraisal theory. He defined the sentimental properties of linguistic assets which is used in sentimental analysis. Along with these, the techniques used in text classification and summarization can also be applied to sentiment analysis. Despite these two strategies utilizing similar logic, sentiment analysis focus on categorizing every review while opinion summarization is on how to effectively obtain the mood expressed in a text and compile them from a high dimensional data set.

A. Development of Linguistic Resource

The appraisal theory depicted in [8] details the issues in opinion mining along with a context of linguistic resource as to how users express their inter-subjective and ideological opinions. Users would mostly pose three characteristics in their development strategy for linguistic properties namely subjectivity, orientation and strength of term attitude. Words like good, excellent and best are positive while bad, wrong and worst are negative. Along with these there are 4 major methods in developing linguistic resources namely;

- Conjunction Method
- Pointwise Mutual Information
- WordNet exploring method
- Gloss Classification method

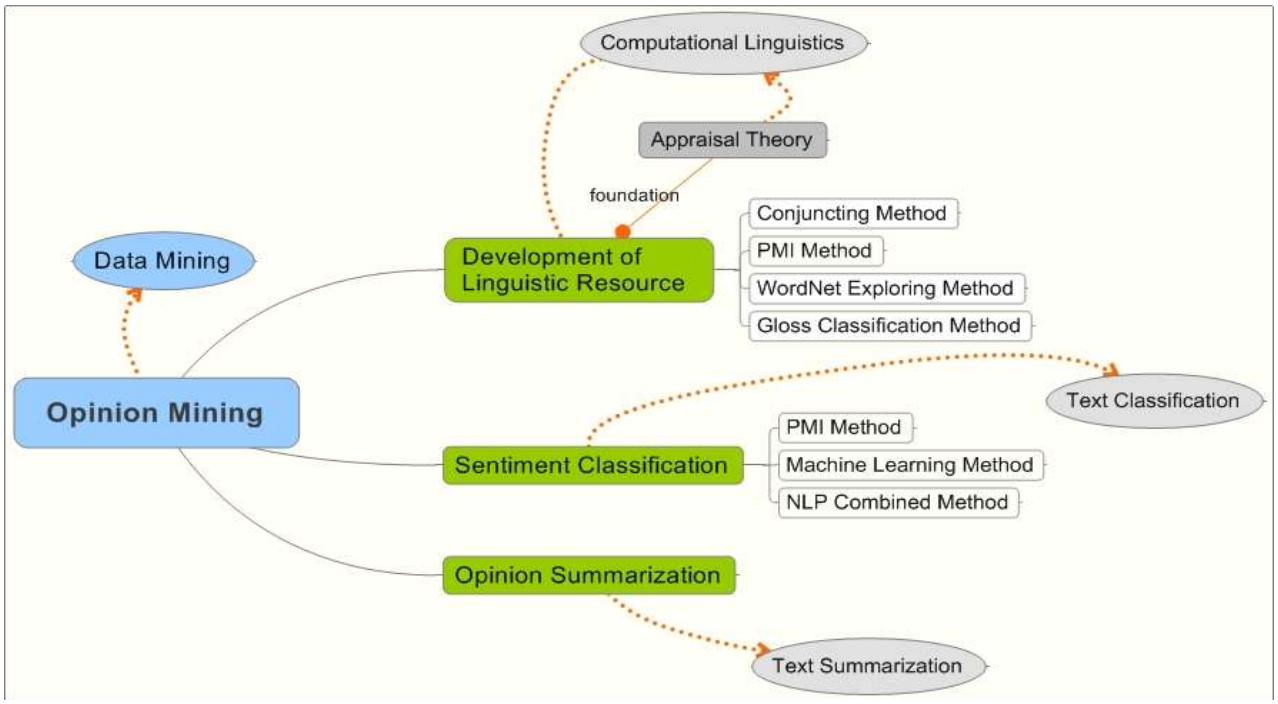
B. Sentiment Classification

Sentiment classification is the process of identifying the sentiment- or polarity - of a piece of text or a document. In this section, wepresent three methods only for classifying reviews as positive ornegative.

- Pointwise Mutual Information
- Machine Learning Methods
- NLP Combined Method

C. Opinion Summarization

Opinion analysis refers to automaticallyextracting attitudes, opinions, evaluations and sentiments from a text. Differing from the earlier text summarization which tries to phrase an effective content with fewer texts, Opinion mining purposesto deliver the complete opinion/sentiment expressed in a huge text. Its moderately insignificant that later is a sub-task of the former. For example, while analyzing product reviews, each and every review is quantified into a proportion of positives and negatives which would indeed be utilized to project the favorableness about the product. We never consider how the overall opinion of each specific feature about a product is summarized. This process is taken care by analyzing several opinion mining systems.



IV. EXPERIMENTAL ANALYSIS

The principal technique of our proposal is the SVM based categorizer. Bo Pang et al [4] depicts clearly as to how SVM has supremacy over all the other supervised learning methods. We utilize the bag of words feature. The WordNet synonymy graph is used to derive the asset or usefulness of an adjective in a good vs bad scenario. These would be utilized like a standard binary value in the feature vectors for SVM. Conversely there is an issue of cacophony being subjected by few texts that might describe the strategy of the movie which was reviewed. To alleviate this dispute we use a subjectivity detector to differentiate the parts that describes the movie from those that describes about the content of the movie. The "about" were utilized in further analysis. Post this classification the probability to asses from the SVM's which would conclude if the review was positive or else negative would be done.

For this to be deployed successfully we need to designate the notion of strengths in the scenario of good vs bad. Fundamental idea of calculating these strengths were developed by Charles Osgood's theory of Semantic Differentiation [9]. WordNet's synonymy graph helps to determine the subjective merits of adjectives. amps et al [10] depicts the evaluation function EVA

EVA(w) = *(d(w,bad)-d(w,good)/ (d(good, bad))*

It's an efficient way of evaluative strength of an adjective. The geodesic function d(wi;wj) is given by the distance between words wiand wjin the *WordNet* synonymy graph. The values are divided by d(good, bad) ,*i.e.*, the distance between the two reference words to restrict the values to [-1,1].

The "about" sentences can be detected in a related method like sentimental analysis with respect to the fact that if we develop a learning process on "about" vs "of" categorization, it can help us detect the "about" sentences in a document. As depicted in [1] we fail on an aspect of info present in structural and semantic relationships among the sentences in a text. To overcome this issue we deploy two types of weightages. The first one depends on individual weights that were determined by SVM's "good" vs "bad" categorization. The second kind is that of mutual weights which is a quantity of the affinity between two sentences to present in the same class in a "about" vs "of" categorization.

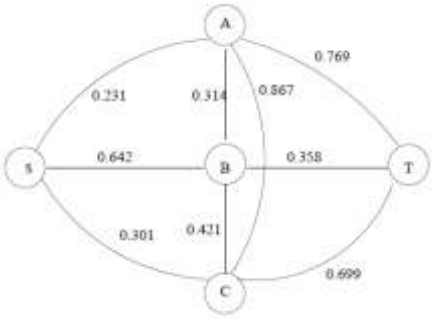
Once we fetch all these weights, we develop each sentence to calculate total adjectival strength which is the sum of the strengths of all the adjectives inthat sentence. The mutual weight is the difference betweenthe weights of the two sentences. This value multiplied bya distance measure and appropriate scaling factors gives thefinal value for the mutual weight between two sentences.Once all the individual and mutual weights have beencomputed, we employ a graph-cut based partition techniqueas described in Bo Pang et al.(2004)[1].

One of the findings in determination of "about" sentences is the mutual relationship that delivers a prodigious pact of treasured info which can meaningfully increase the accuracy of prediction. A fundamental query that arises is the fact as to why these mutual relationships aren't used to increase the accuracy of opinion mining. Initial assignment was to begin with appropriate metrics for resemblance between texts since feature vectors might contain a significant portion of the info about the content of the text. Number of common features in every pair of vector is documented and are scaled down to [0,1]. These are denoted as mutual similarity co efficient between a pair of documents. They are indeed given by;

$$MSC(d_i, d_j) = \frac{\sum_k (F_i(f_k) * F_j(f_k)) - s_{min}}{s_{max} - s_{min}}$$

Where f_k is the k^{th} feature.
$F_i(f_k)$ is a function that takes the value 1 if the k^{th} featureis present in the i^{th} document and is 0 otherwises_{max} is the largest value of the number of common features between any two documentss_{min} is the smallest value of the number of common featuresbetween any two documents.

SVM was also trained so as to detect the chances of the review between positive and negative moderately than the category label [11]. These probabilities along with the mutual similarity co efficient give rise to a weight matrix where we can deploy the graph cut partition technique as discussed in [1]. The source and the sink refers to positive and negative moods expressed respectively. Theedges linking a document to source have their capacity asthe likelihood of the text being a positive one. Likewisewe assign edge dimensions for the edges to sink. Theedges between documents have the same capacity as theMSC of the two documents. For example, consider a setof 3 documents with values as shown in Table 1. Then theedge weights in the minimum cut setup would be as indicatedin Figure below.



V. EXPERIMENTAL RESULTS AND FINDINGS

The movie corpus used for this assignment was the tagged corpus introduced by Bo Pang et al in [1]. It contains 1000 positive and negative reviews. It was completely run through a POS tagger as the POS tags were required for some later tasks. The tagger utilizedwas Stanford Log-Linear Model Tagger v1.04. The correspondingdocuments were used for partiality detection.This module has twoparts. The first was for the valuation of discrete weights.Here we used an SVM that was accomplished to forecastlikelihoodapproximationsmoderately than class

labels. The dataset used wasthe Subjectivity Dataset, introduced by Bo Pang et al. [1]. This corpus contains 5000 movie-review snippetsand 5000 plot summaries. The SVM package used waslibsvm-2.71[12].

The second module was to calculate the mutual weights.Here we used the Ford-Fullkerson algorithm to attain theminimum-cut. For calculation of the mutual weights weexperimented with a number of measures. Consider twosentences d lines apart. Let w_x and w_y be the weights ofthe sentences as obtained by summing the strengths of allthe adjectives in the sentences x and y resp. For example,if x is the sentence"The movie was excellent with outstandingperformances from all actors", then wx would bew$_{excellent}$+w$_{outstanding}$.
Let assoc(si; sj) be the mutual weight for the sentence pair siand sj. Then we have;

$$assoc(s_i, sj) = c * f(d) * g(w_i, w_j)$$

c is just a constant factor. A larger value of c implies thatthe algorithm will be more loath at putting sentences not having a great deal of similarity in different classes. Differentvalues of c were tried out with an aim of optimizing theclassification results downstream.

As an initial step we ran our tests on the complete document without considering any extracts of "about" sentences. Many features were considered. The first assumption was using adjectives as a feature since it plays a vital role in determining the polarity of a document. Secondly BNS feature selection algorithm was utilized. In the primary approach the adjective weights were considered and was multiplied by any appropriate multiplier if any modifiers were present. In case if the adjective was found between a "not" and a punctuation mark, then the weightage was multiplied by -1. All these values were negated from every document. Words like "good", "very good" and "not good" were all considered to be the instance of same feature "good". The weightage for these would be w$_{good}$, m$_{very}$* w$_{good}$and -1* w$_{good}$.Where m$_x$ denotes the weight of the modifier x. Using thesefeature vectors we obtained a five-fold cross validation accuracyof 68.1% over the dataset.

We then took the top 32000 unigrams as our features.The adjectives in these were sorted out for a separate treatment.For other types we used just binary values,1if the feature is present and 0 if absent. Foradjectives, we tried out the same methodology as earlier. Thefive-fold cross authentication accurateness in this case was found tobe 70.2%.

A finding was that feature like "better" looks both like an adjective and a non-adjective. They were all prefixed with "ADJ". We also prefixed "NOT" for every feature occurring between "not" and a punctuation mark. Accuracy this way was found to be 68.29%. adjectives with a negative occurrence was labelled as "NEG ADJ" tag and the positive one was labelled as "POS ADJ". With these features the accuracy droppeddown to 65.5%.

At last we considered 32000 unigrams and separated the adjectives as earlier. The weight of the adjectives in the feature vector of the document was utilized. For other features still the binary values were used. Accuracy was found to be 75.8% with these changes in place. Since it gave best results so far we proceeded with this approach for further experiments from the documents that were obtained after the detection of the "about" sentences. We initially chose f(d) = 1/d^{2}and g(w$_i$;w$_j$) =|jw$_i$-w$_j$ |. Different values were substituted for c. for c=10 accuracy was 70%. For c=100 accuracy was 67.5%.

This is in pact with the perception that a lower valuefor c should produce better results. To additionallyauthorize this theorywe tried c = 1. For this case the accuracy was 67.5%.This possibly shows that too strict a consequence for a variationor distance between sentences also leads to a dropin accuracy. But a decline of accuracy on use of"about"extracts was pawn to the prospects. This is perhapsdue to the crude task that was utilized to classify sentencesimilarities. A better measure can be expected to give betterresults. We chose such a simple function because the complicationinvolved in the calculation of statistically reliable roles like the Mutual Information Quotient appeared tobe unaffordable in this case.

We then experimented by using distance and contextualsimilarity in isolation. With just a distance measure andc = 100, we obtained an accuracy of 65.8%. In the samecase, using c = 10 gave an accuracy of 68%.Using just the contextual similarity measure gave an accuracyof 68% both for c = 10 and c = 100.Till this point we hadn't taken into account the fact thatusing the mutual similarities between the documents can beused to find out the problems with current predicted labelsand can thus provide a significant increase in accuracy. Wedecided to apply this technique described in Sec. 3.3 to theresults obtained from all the previous steps.

For complete documents using weights for adjectives and binary values for other features, application of this techniqueimproved the accuracy to an overwhelming 95.6%.All the results before and after the application of this technique are listed in Table. We also tried out the use ofBNS feature selection algorithm but no significant changein results was observed.

S no	Type of documents	Before Graph-cut	After Graph-cut
1	Full documents	75.80%	95.60%
2	"about" extracts with distance and context info c = 100	65.65%	94.20%

3	”about” extracts with distance and context info c = 10	70%	92%
4	”about” extracts with distance and context info c = 1	67%	93.50%
5	”about” extracts with distance info c = 100	65.80%	91%
6	”about” extracts with distance info c = 10	68%	89.40%
7	”about” extracts with context info c = 100	68%	84.20%
8	”about” extracts with context info c = 10	68%	84%

VI. CONCLUSION

Clearly, the main strength of our approach lies in showinghow strong influence mutual relationships betweendocuments can have on their sentiment analysis. The wayin which we have used the graph-cut technique for this taskprovides a very simple yet efficient framework for incorporatingthis information. Moreover, this technique can beapplied to improve the accuracy of predictions in any classificationtask over a set of test documents.

Finally, opinion summarization methods were examined whichinclude feature extraction, sentiment assignment, andvisualization. Feature extraction and sentiment assignment aresubtasks of feature-level sentiment classification whilevisualization is about the effective presentation of the summarizedopinion.

REFERENCES

[1] Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity SummarizationBased on Minimum Cuts, Proceedings of ACL, 2004.
[2] https://en.wikipedia.org/wiki/Bag-of-words_model_in_computer_vision
[3] PeterTurney. 2002. Thumbs up or thumbs down? Semanticorientation applied to unsupervised classication of reviews. In Proc. of the ACL.
[4] Bo Pang, Lillian Lee and ShivakumarVaithyanathan,Thumbs Up? Sentiment Classification Using MachineLearning Techniques, Proceedings of EMNLP 2002, pp79-86.
[5] T. Nasukawa and J. Yi, ―Sentiment analysis: Capturing favorabilityusing natural language processing‖, in Proceedings of theConference on Knowledge Capture (K-CAP), 2003.
[6] S. Das and M. Chen, ―Yahoo! for Amazon: Extracting marketsentiment from stock message boards‖, in Proceedings of the AsiaPacific Finance Association annual Conference (APFA), 2001.
[7] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack,―Sentimentanalyzer: Extracting sentiments about a given topic using naturallanguage processing techniques‖, in Proceedings of the IEEEInternational Conference on Data Mining (ICDM), 2003.
[8] Martin, J. R. and White, P. R. R. The Language of Evaluation: Appraisal in English. Palgrave, London, 2005.
[9] Osgood, C. E., G. J. Succi, and P. H. Tannenbaum, 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana IL.
[10] JaapKamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke.*Using WordNet to measure semantic orientation of adjectives.*In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), volume IV, pages 1115-1118. European Language Resources Association, Paris, 2004.
[11] T.-F. Wu, C.-J.Lin, and R. C. Weng. *Probability estimates for multi-class classification by pairwise coupling.* In S. Thrun, L. Saul, and B. Scholkopf, editors, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2004. URL http://www.csie.ntu.edu.tw/ cjlin/papers/svmprob.pdf.h
[12] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm