A Review on Movie Script Classification using Sentimental Analysis Approach

Naisargi Sureja
Research Scholar
Computer Engineering Department, Darshan Engineering College,Rajkot,India

***Abstract* – Sentimental analysis of a movie review plays an important role in understanding the sentiment conveyed by the user towards the movie. Now a day's analysis is more towards specification of categorization or classification of movies in user specific choice which are dependent in either the mood or emotion or choice of user perspective. So proposed a method where movies based on subtitles that are played by characters contains such kind of sentimental words in sentences so that movie can be classified into specific genre. In current work, using the subtitles in dataset and considering movie genre like comedy, thriller, action drama and horror, it can be developed sentimental analysis model using lexicons that are context specific to each genre under consideration. Proposed methodology will play important role in field of movies industries too.**

***IndexTerms* - Sentiment analysis, Movie genres, Lexicon, SVM, SO-CAL, Naïve Bayes**

I. INTRODUCTION

Sentiment analysis [1] is a methodology by which we find out the sentimental orientation of a piece of text. Using it, we can infer whether a particular person has conveyed a positive or negative sentiment in the said text under consideration. We tackled the issue of aspect based sentiment analysis of movie reviews in our previous publication [2]. In it we use the concept of "driving factors", which enhanced the overall classification accuracy by amplifying the effect of certain movie aspects with respect to others. In the current work, we tend to use the same concept, but for reviews with different genre. Many researchers have done work on Aspect based analysis of review, be it movie or customer review. Also many algorithms have been developed for the same. But not much work has been done on genre specific aspect based analysis. Genre specific reviews demand special techniques while analysing as such reviews contain sentences or words that have unique meaning based on the context i.e. genre in which they are used. Thus in this paper we try to develop an unsupervised aspect based analysis model that uses context i.e. genre specific lexicons. We made use of separate lexicons for each genre, and using this we try to inculcate context sensitivity into the model. Also using aspect based analysis, we try to develop fine grained aspect level analysis model.

Bo Pang et al [3] suggested many analysis methods including supervised models like Naïve Bayes, Support Vector Machines and Maximum Entropy classifiers. Using different features like unigrams, unigrams and bigrams, top unigrams, adjectives, they carried out experiments and compared their results. Their work gave researchers great insight into the field of sentiment analysis. Research conducted by Abd. Samad Hasan Basari et al [4] increased the overall accuracy of analysis models like Support Vector Machines by coupling them with Particle Swarm Optimization algorithm. Their work increased the overall accuracy obtained by them from 71.87% to 77%. Recent years have shown an increase in aspect based sentiment analysis as better analysis of review can be done if polarities of individual aspects are considered. Tun Thura Thet et al [5] developed a model for fine-grained sentiment orientation and sentiment strength analysis of various aspects of the movie. They formulated domain and generic opinion lexicons to score words in the review. The individual word score obtained via these lexicons was propagated over the entire document by using the inter-word dependencies obtained through the usage of dependency trees. A feature based heuristic model for aspect level sentiment analysis was developed by V.K. Singh et al [6]. In their scheme they identified various movie aspects in the review text and assigned labels to them. Each aspect text is then scored using SentiWordNet [7] with feature selection comprising of adverbs, adjectives, n-grams and verb features. They obtained the overall document score by averaging individual aspect score. Maite Taboada et al [8] in their work Genre-Based Paragraph Classification for Sentiment Analysis distinguished between different types of paragraphs of movie reviews using a taxonomy and classification system.

II. RELATED WORK

For Sentimental Analysis there are various lexical resources in use. We discuss Dictionary and SentiWord-Net in this section.

Dictionary

All sentiment analysis tools require a list of words or phrases with positive and negative connotation, and such a list of words is referred to as a dictionary. Dictionary is an important lexical resource for Sentiment Analysis.
A single dictionary for all the domains, is difficult to generate. This happens because of the domain specificity of words. Certain words convey different sentiments in different domains. For example:

- Word like fingerprints" conveys a major breakthrough in a criminal investigation whereas it will be negative for smartphone manufacturers.
- Freezing" is good for a refrigerator but pretty bad for software applications.

- We want the movie to be unpredictable" but not our cell phones.

A few popular dictionaries are discussed in the following sections.

The Loughran and McDonald Financial Sentiment Dictionary:
Loughran and McDonald (2011) show how, applying a general sentiment word list to accounting and financial topics, will lead to a high rate of misclassification. They found that around three-fourths of the negative word in a general sentiment dictionary were not negative in the financial domain. Sothey created the dictionary, The Loughran and McDonald Financial Sentiment Dictionary". It is a publicly available domain-specific dictionary and it contains custom lists of positive and negative words specific to the accounting and financial domain.

Lexicoder Sentiment Dictionary (LSD):

Lexicoder Sentiment Dictionary (LSD) is also a domain-specific dictionary. It expands the score of coverage of existing sentiment dictionaries, by removing neutral and ambiguous words and then extracting the most frequent ones. Some important features of this dictionary are the implementation of basic word sense disambiguation with the use of phrases, truncation and preprocessing, as well as the effort to deal with negations.

WordStat Sentiment Dictionary:

The WordStat Sentiment Dictionary was formed by combining words from the Harvard IV dictionary, the Regressive Imagery dictionary (Martindale, 2003) and the Linguistic and Word Count dictionary (Pennebaker, 2007). It contains a list of more than 4733 negative and 2428 positive word patterns. Sentiment is not predicted by these word patterns but by a set of rules that take into account negations.

Sentiwordnet

SentiwordNet is a lexical resource in which each wordnet synset `s' is associated to three numerical scores Obj(s), Pos(s) and Neg(s), which describe how objective, positive and negative the terms contained in the synset are. Each of the three scores range from 0.0 to 1.0, and their sum is 1.0 for each synset. A graded evaluation of opinion, as opposed to hard evaluation, proves to be helpful in the development of opinion mining applications.

III. DETERMINING MOVIE GENRES TO BE USED

IMDB recognizes a total of 27 different genres. However, it has been observed by [5] that some of these genres show a high correlation. For example, movies which have been tagged as Mystery are very likely to have the genre Thriller associated with it as well. This was also verified by our clustering methods. This helps us in reducing the number of clusters. On performing clustering algorithms on closely related genres such as Mystery and Thriller, it has been observed by [1],[2],[3],[4] and [5] that there is a very thin line of difference between two such very closely related genres and currently, no research describes decisive quantitative factors using which the two genres can be distinguished. Of course, Mystery/Thriller combination is just one example. [1],[2],[3] and [4] have classified the movies into Comedies, Horror, Drama and Action. All the other genres, such as Thriller or Crime, would fall in either one of these categories. [5] has clubbed the genres into five groups using hierarchical clustering as follows:

Cluster	Genres Included
Cluster 1	Short, Drama Comedy, Romance, Family, Music, Fantasy, Sport, Musical
Cluster 2	Thriller, Horror, Action, Crime, Adventure, Sci-Fi, Mystery, Animation, Western
Cluster 3	Documentary, History Biography, War, News
Cluster 4	Reality-TV, Game Show, Talk Show
Cluster 5	Adult

Table I.

The above classification as shown in Table I includes some genres such as Reality-TV, Game Show, Talk Show, Adult, News etc. Our corpus did not contain scripts belonging to these genres. Hence, based on the linkage matrices provided in [5], we created a classification more suitable for our study.

Cluster	Genres Included
Cluster 1	Drama, Comedy, Romance, Family, Sport, Musical
Cluster 2	Action, Western, War
Cluster 3	Sci-Fi, Adventure, Fantasy, Animation
Cluster 4	Crime, Mystery, Thriller
Cluster 5	Horror

Table II.

The above classification as shown in Table II was derived by grouping together movie genres which show high similarity in terms of co-occurrence and repetition of movie keywords as defined by IMDB. For more details regarding how the keywords were used to determine the group of genres, please refer [5].

IV. PROPOSED METHODS

The method aims at developing a lexicon based aspect oriented analysis approach for genre specific reviews. Fig.1 describes the flow of the proposed method.

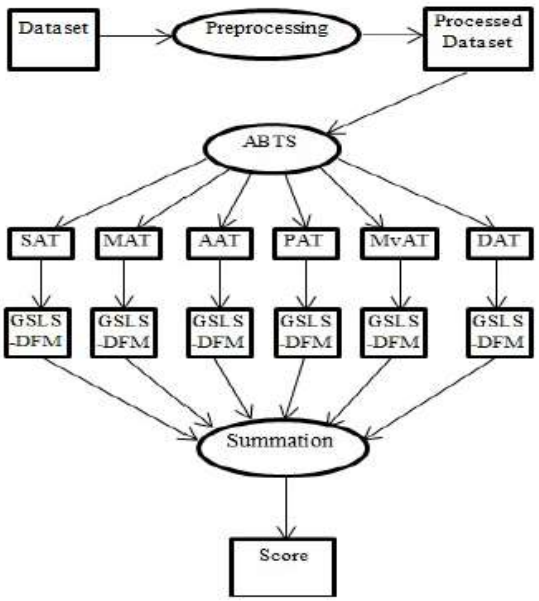

Figure 1. Proposed System Architecture

ABTS= Aspect Based Text Separator, SAT = Screenplay Aspect Text, MAT = Music Aspect Text, AAT = Acting Aspect Text, PAT = Plot Aspect Text, MvAT = Movie Aspect Text, DAT = Direction Aspect Text, GSLS-DFM = Genre Specific Lexicon based Scoring and Driving Factor Multiplication.

We used the dataset that Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith [11] used in their experiments. The dataset was in XML format and each file contained movie details like name of the movie, genre of the movie, date of release and also sited web links for obtaining full reviews. Pre-processing was required for the dataset as it was not in accordance with our requirements. We extracted the links and the genre of each movie from the dataset. Then from the corresponding links we acquired the full movie review manually. However the dataset didn't contain ratings for the movies. We needed the movie ratings so that we may compare the classification obtained using our methodology with the actual movie rating and determine the accuracy of our classification. We used movielens dataset [12] for determining the movie rating. Having done so, we prepared a preprocessed dataset containing information like the name of the movie, its genre, its rating and the review text. This dataset was then used further for experimental purposes.

Now the next step was that of separating the review text into aspect specific text. As mentioned in our previous publication [2] we tend to use Aspect Based Text Separator (ABTS) for this purpose. ABTS separates the review text into different groups based on the movie aspects. It does this using an aspect lexicon. The process of ABTS separation and solution to problems like ambiguity are explained thoroughly in the previous publication. The next step was that of the classification of these separated aspect texts. As mentioned previously, sometimes certain words tend to have different meaning based on the context in which they are used. For example take the word 'funny'. It tends to denote a positive orientation if used in the review of a comedy movie, as one would expect a comedy movie to be funny. But if used in a movie of horror genre, then the orientation of the word becomes somewhat negative as horror movies tend not to be funny. People, in general, tend to enjoy horror movies that scare them and which are thrilling. If someone uses the word funny to describe a horror movie, then there is a very good chance that the word was used in a negative sense. Thus to account for all these context sensitive words, we tend to develop a genre specific lexicon.

This lexicon would contain certain words whose orientation would depend on the genre in which they are used. We formulated a list of top 500 frequently used adjectives in everyday life and formed a lexicon out of these words. Now we wanted to assign orientation to these words based on the movie genre. Thus we used the methodology of Semantic orientation to do so.
Before we define what Semantic Orientation (SO) [13] is, let us first define Pointwise Mutual Information (PMI). PMI between two words is the amount of information that we acquire about the presence of one word when we observe the other [14]. The formula for PMI is:

$$PMI(word_1, word_2) = \log_2\left(\frac{p(word_1 \& word_2)}{p(word_1)p(word_2)}\right)$$

Here *p(word1 & word2)* is the probability of *word1* and *word2* occurring together. If the words are statistically independent, then the probability of their co-occurrence is given by *p(word1)p(word2)*. The degree of statistical dependence between words is given by the ratio of *p(word1 & word2)* and *p(word1)p(word2)*. The semantic orientation of a word, *word* is calculated by:

$$SO(word) = PMI(word, X) - PMI(word, Y)$$

Here X denotes a positively oriented word or string of words and Y denotes a negatively oriented word or a string of words. Thus we find the co-occurrence of *word* with a positive word and with a negative word. Then we subtract the PMI obtained with the positive and negative word to get the overall orientation of the word. Thus if we obtain a negative value, the overall SO is negative, and it means that the word under consideration occurs more closely with the negative word string and similarly if the result the positive, the word is closely associated with the positive word string. Thus the SO methodology can be easily used for preparing genre specific lexicon. Now the question that remained was that of finding the co-occurrence of words. Previously researchers used to use AltaVista search engine and it's NEAR operator to do so [14]. But since AltaVista is no longer operational, we decided to query a movie dataset instead of the internet to find out the co-occurrence of words. The query which we issued over the dataset returned us with the hitcount of our search query. This hitcount was then used to calculate the probability for finding the co-occurrence. We used the Large Movie Review Dataset [15] for the same. The NEAR operator functionality was programmatically recreated using Boolean operators to work on our dataset. Two find co-occurrence of two words, *word1* and *word2* we issued a Boolean query over our dataset as: "word1 word2" OR "word2 word1" OR "word1 * word2" OR "word2 * word1" [16]. The above query considers all the cases related to the positioning of the words. Here '*' represents a wildcard, which means there can be single or multiple words between the two words. As mentioned earlier we calculated the proximity of the adjective towards both positive and negative words or strings, via acquired hit count. We prepared a positive and negative string for each genre under consideration. Consider the genre 'action'. For this genre we prepared a positive string as follows "action AND good" and negative string as "action AND bad". When the adjective was queried along with this string, the results returned the documents in which the words 'action' and 'good' co-occurred along with the adjective. Similarly strings were created for other genres like comedy, crime, drama and horror. We worked on the assumption that in the reviews, the user might have mentioned to which genre the movie belonged to. It was a fair assumption according to us as most of the reviews we encountered had some indication contained in them regarding the genre of the movie.

V. PROPOSED ALGORITHM

We outline our semantic orientation calculator, SO-CAL. SO-CAL extracts words from a text, and aggregates their semantic orientation value, which is in turn extracted from a set of dictionaries. SO-CAL uses five dictionaries: Four lexical dictionaries with 2,257 adjectives, 1,142 nouns, 903 verbs, and 745 adverbs, and a fifth dictionary containing 177 intensifying expressions. Although the majority of the entries are single words, the calculator also allows for multiword entries written in regular expression like language.
The SO-carrying words in these dictionaries were taken from a variety of sources, the three largest a corpus of 400 reviews from opinions.com, first used by Taboada and Grieve (2004), a 100 text subset of the 2,000 movie reviews in the Polarity Dataset (Pang and Lee, 2004), and words from the General Inquirer dictionary (Stone, 1997). Each of the open-class words were given a hand-ranked SO value between 5 and -5 (neutral or zero-value words are not included in the dictionary) by a native English speaker. The numerical values were chosen to reflect both the prior polarity and strength of the word, averaged across likely interpretations. For example, the word *phenomenal* is a 5, *nicely* a 2, *disgust* a -3, and *monstrosity* a -5. The dictionary was later reviewed by a committee of three other researchers in order to minimize the subjectivity of ranking SO by hand.
Our calculator moves beyond simple averaging of each word's semantic orientation value, and implements and expands on the insights of Polanyi and Zaenen(2006) with respect to contextual valence shifters. We implement negation by shifting the SO value of a word towards the opposite polarity (*not terrible*, for instance, is calculated as -5+4 = -1). Intensification is modelled using percentage modifiers (*very engaging*: 4x125% = 5). We also ignore words appearing within the scope of *irrealis* markers such as certain verbs, modals, and punctuation, and decrease the weight of words which appear often in the text. In order to counter positive linguistic 65 bias (Boucher and Osgood, 1969), a problem for lexicon-based sentiment classifiers (Kennedy and Inkpen, 2006), we increase the final SO of any negative expression appearing in the text.
The performance of SO-CAL tends to be in the 76-81% range. We have tested on informal movie, book and product reviews and on the Polarity Dataset (Pang and Lee, 2004). The performance on movie reviews tends to be on the lower end of the scale. Our baseline for movies, described in Section 5, is 77.7%. We believe that we have reached a ceiling in terms of word- and phrase-level performance, and most future improvements need to come from discourse features. The stage classification described in this paper is one of them.

V.IMPLEMENTATION & RESULTS

The dataset was acquired from the Large movie review dataset site of Stanford AI Lab [14][15].The dataset was collected from IMDB and contains around 50,000 reviews, out of which 25000 are positive and 25000 are negative. Though there is no specific time span for review collection from IMDB, but it was ensured that no more than 30 reviews from a single movie get included in the final dataset. Because of even number of positive and negative reviews, so the minimum accuracy that we can obtain from the experiment is 50%. The dataset contains only highly positive and highly negative reviews. The authors of the dataset included a negative review only if it scored ≤ 4 out of 10 and included a positive review if it scored ≥ 7 out of 10 on a benchmark set by them [14]. Neutral reviews were omitted. It was seen that since the size of the reviews was varying and also since each aspect was commented on, in unequal number of sentences, the aspect based text separator separated the review into various aspects having unequal text distribution. It was also observed that in some of the reviews, all the aspects were not commented on. The score of the classifier for these cases were made 0. As mentioned in the previous section a Naïve Bayes classifier was used for classifying the separated aspect based text. The aspect based text given as input to individual classifiers was divided into training and testing data with a ratio of 70:30. The experiment was conducted with 1000 iterations, and during each iteration the driving factors were assigned random values following the constraints of (1). The driving factors which gave the highest accuracy were chosen as the best driving factors for the particular dataset under consideration.

Suppose we have a review X and it contains user's opinion about two factors F1 and F2. Also the overall orientation of the review is positive in nature. The user has given a positive review about F1 and a negative about F2. Also the amount of text in the review for F1 aspect is less as compared to the F2 aspect. Now if we use any non-aspect based sentiment analysis method then since text size of F2 is greater than text size ofF1 and also since F2 is negative in orientation, the overall document score will tend to reduce and skew towards negativity. On the other hand if driving factors are used and F1 is given more importance the document score will better reflect the positivity of the review. Since each aspect of a movie is analysed separately in this method, we can track the effect each aspect has towards the overall score of the document. This individual aspect based tracking can be used in a fined grained aspect based recommendation system, which recommends movies based on its various aspects instead of the overall rating of the movie. Also this method can be applied on a product review dataset thus enabling us to see what opinion each user has on the various aspects of the product, thus helping in the development of proper product placement strategy. It is very difficult to acquire such in-depth knowledge from the dataset using non-aspect based methods.

The various performance measures used were [6]:
Accuracy = (Total correctly classified documents / Total number of documents)
Precision = tp / (tp +fp)
Specificity = (tn / Total number of negatively oriented documents in the dataset)
Recall = (tp / Total number of positively oriented documents in the dataset)
Where tp, fp and tn are the true positives, false positives and true negatives obtained during the classification.

VI. CONCLUSION

Using mentioned methodology we got the results that for action genre we got plot, movie and direction as the most important factors as these aspects had the highest values, for comedy we got acting, plot and movie, for crime we got screenplay, music and plot, for drama we got music, movie and direction and for horror we got music, direction and movie. These results obtained are only for the particular dataset under consideration. Other datasets may give different results, but by plain observation of the results one can say that they are somewhat correct. Take the example of comedy genre. In this genre people like the way actor act and make them laugh. Also the storyline and hilarious plot of the movie might make someone crack a smile. Thus these things are evident from the results we got for the comedy genre. Using the driving factors we were able to extract the most importance aspects for a particular dataset under consideration. Thus by using this methodology, we can identify importance aspects across various datasets and across various genres.

REFERENCES

[1] Sentiment Analysis, Wikipedia – The Free Encyclopedia (http://en.wikipedia.org/wiki/Sentiment_analysis).
[2] Viraj Parkhe, Bhaskar Biswas, "Aspect Based Sentiment Analysis of Movie Reviews: Finding the polarity directing aspects", International Conference on Soft Computing and Machine Intelligence, New Delhi, India, 2014.
[3] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine LearningTechniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.
[4] Abd. Samad Hasan Basaria, Burairah Hussina, I. Gede Pramudya Anantaa, Junta Zeniarjab, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization", in Malaysian Technical Universities Conference on Engineering & Technology 2012, MUCET 2012 Part 4 – Information And Communication Technology.
[5] Tun Thura Thet, Jin-Cheon Na and Christopher S.G. Khoo, "Aspectbased sentiment analysis of movie reviews on discussion boards"., in Journal of Information Science,2010 36:823.
[6] V.K. Singh, R. Piryani, A. Uddin and P. Waila, "Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification", Proceedings of the 2013 International Muli Conference on Automation, Communication, Computing, Control and Compressed Sensing, Kerala-India, March 2013, IEEE Xplore, pp. 712-717.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[7] SentiWordNet, lexical resource for opinion mining,(http://sentiwordnet.isti.cnr.it/).

[8] Maite Taboada, Julian Brook, and Manfred Stede ,"Genre-Based Paragraph Classification for Sentiment Analysis" ,Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 62–70,Queen Mary University of London, September 2009.
[9] H. Zhou, T. Hermans, A. V. Karandikar, J. M. Rehg, "Movie Genre Classification via Scene Categorization", Proc. 10th international conference on Multimedia, pp. 747-750, 2010.
[10] Z. Rasheed and M. Shah, "Movie genre classification by exploiting audio-visual features of previews", Proc. the 16th International Conference on Pattern Recognition vol.2, no., pp. 1086- 1089 vol.2, 2002.
[11] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," IEEE Transactions on Circuits and Systems for Video Technology, vol.15, no.1, pp. 52- 64, Jan. 2005.
[12] A. Austin, E. Moore, U. Gupta, and P. Chordia, "Characterization of movie genre based on music score," IEEE International Conference on Acoustics Speech and Signal Processing, pp.421-424, 2010
[13] H. Bulut and S. Korukoglu, "Analysis and Clustering of Movie Genres", Journal Of Computing, Volume 3, Issue 10, pp.16-23, October 2011
[14] Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher, "Learning Word Vectors for Sentiment Analysis", in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2011.
[15] Large Movie Review Dataset, Acquired from Stanford AI Lab: (http://ai.stanford.edu/~amaas/data/sentiment).