

# Using Sentimental Analysis Approach Review on Classification of Movie Script

<sup>1</sup>Naisargi D. Sureja, <sup>2</sup>Prof. Firoz A. Sherasiya

<sup>1</sup>Research Scholar, Department of Computer Engineering, Darshan Institute of Engineering and Technology Rajkot, Gujarat

<sup>2</sup>Assistant Professor, Department of Computer Engineering, Darshan Institute of Engineering and Technology Rajkot, Gujarat

**Abstract**—Now a day's sentimental analysis has become one of the greatest innovation and efficient technique of information and data analysis. 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.

**Index Terms** - : Sentiment analysis, Movie genres, Lexicon, SentiWordNet, WordNet

## I. INTRODUCTION

Using methodology of Sentimental analysis <sup>[1]</sup>we can find out the sentimental orientation of a piece of text. We tackled the issue of aspect based sentiment analysis of movie reviews <sup>[2]</sup>. Genre specific reviews demand special techniques while analyzing as such reviews contain sentences or words that have unique meaning based on the context i.e. genre in which they are used. 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.

Movie genre classification is a challenging problem with many potential applications. Whereas many prior approaches rely on image, audio, or motion features to classify movies, we consider using textual content analysis instead, which is a comparatively less computationally expensive and time consuming process. For example if we find genre based Bollywood movie than it gives results based on actor, actress, music, etc. If we find particular name of genre than it gives theoretical information over there but it is not giving particular list of movie. In this light, we focus on sentimental analysis of movies from its scripts.

## II. RELATED WORK

IMDB recognizes a total of 27 different genres. However, it has been observed by <sup>[5]</sup> 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 <sup>[1][2][3]</sup> 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.

For Sentimental Analysis there are various lexical resources in use. We discuss Dictionary and SentiWordNet<sup>[7]</sup> 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.

#### *Lexicoder Sentiment Dictionary (LSD):*

Lexicoder Sentiment Dictionary (LSD) is also a domain-specific dictionary. It expands the scope 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<sup>[7]</sup>.

#### *WordNet:*

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing<sup>[8]</sup>.

### III. PROPOSED METHODOLOGY

Proposed system can be used in any application that requires fast and easily adaptable text categorization in terms of stylistically homogeneous categories. The sentiment analyzer is basically the main part of proposed framework. It uses general resources like SentiWordNet<sup>[7]</sup>, WordNet<sup>[8]</sup> and any domain specific resource to extract the sentiment-oriented words from each sentence by using the relationship information of (dependencies within) the sentence. The Sentiment Analyzer has two sub modules, which help in calculating the polarity of sentences and documents. The Polarity Calculator (PC) calculates the polarity of a sentence and assigns a score. In order to calculate polarity, PC uses SentiWordNet to identify the positive and negative words and their values assigned by the SentiWordNet. In this process, PC collects the synonyms of a word if it's not found in SentiWordNet. The PC first uses WordNet to get the synonyms.

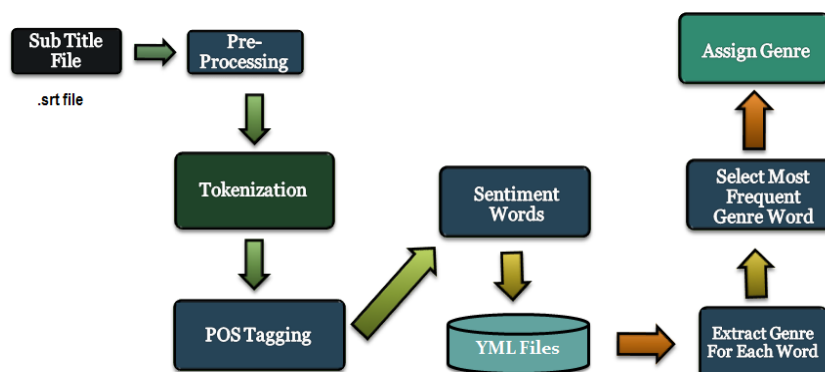


Figure 1. Proposed system Architecture

**Proposed system works as following:**Step 1:

Input Script File of Movie is given to the system,

Step 2:

At next, document of script will perform pre-processing which will follow by cleaning, stop word removal and at the end Tokenization is been performed.

Step 3:

Algorithm Of POS Tagging will parse the whole script and generate part of speech tagging system to every word that is been generated.

Step 4:

Now that every word is been matched with YML files that contains particular movie genre wise YML files, so that all the words can be matched with that file and assign every word a particular sentiment.

Step 5:

Overall at the end all words are been mined and then words with most frequent genre are been selected as that category/genre of particular movie script.

**Pre-processing includes main three parts like:**Sentence Splitting:

In sentence splitting just split paragraph into sentence from end words.

Tokenization:

Tokenization method divides the text of a sentence into sequence of tokens and creates results in tokens consisting of one single word (unigram).

POS tagging:

Last is the process of Part-Of-Speech tagging (POS) allows to automatically tag each word of text in terms of which part of speech it belongs to:

Adverb, noun, pronoun, adjective, verb, interjection, intensifier etc.

**IV. EVALUATION**

Based on precision recall value, it is shown that the accuracy of Action genre is good, that followed by filmy, romance and others. The least precision is of Crime subtitle based movies.

The various performance measures used were <sup>[6]</sup>:

Accuracy = (Total correctly classified documents / Total number of documents)

Precision =  $tp / (tp + fp)$

Specificity =  $(tn / \text{Total number of negatively oriented documents in the dataset})$

Recall =  $(tp / \text{Total number of positively oriented documents in the dataset})$

Where tp, fp and tn are the true positives, false positives and true negatives obtained during the classification.

Genre Name	Precision	Recall	F-Square
Action	0.902439024	0.108108	0.794331
Mystery	0.4	0.6	0.666667
Drama	0.666666667	0.5	0.166667
Romance	0.842105263	0.1875	4.491228
Comedy	0.696428571	0.435897	1.597689
Thriller	0.693877551	0.441176	0.252701

<b>Crime</b>	0.22222222	0	0.222222
<b>Musical</b>	0.75	0.333333	0.416667
<b>Horror</b>	0.818181818	0.194444	0.623737
<b>Filmy</b>	0.893333333	0.119403	0.77393

Table 1. Genre wise accuracy

## V. CONCLUSION AND FUTURE WORK

From Literature Survey conclude that the dissertation target, to mainly on proposed system and we can conclude that the SentiScript will be useful as an application level and commercial level too, to analyze whole script and categorize movie through script conversation with target of faster execution. Overall implementation leads to proposed system framework which categorizes movie script as per genre. By value of precision, recall of each genre, it is shown that Action movie subtitles are properly classified while as crime movie genres are having less accuracy others are having on an average good accuracy value. Thus, it will be helpful not only at research perspective but also over the real time application level which will be helpful even to normal user who want a specific targeted genre based movies.

## VI. FUTURE WORK

The proposed system can be enhanced more by using more genre classification system and can be tried with other aspect feature analysis of subtitles. More over in proposed system have used with English subtitles, other cross platform languages are also be used to get wide area coverage with other languages movie subtitles. In future it can also be used with hybrid classification techniques and can compare results.

## REFERENCES

- [1] Sentiment Analysis, Wikipedia – The Free Encyclopedia([http://en.wikipedia.org/wiki/Sentiment\\_analysis](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 Learning Techniques”, 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, “Aspect based 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 MultiConference 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] Wordnet, A lexical database for English. (<https://wordnet.princeton.edu/>)
- [9] 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.
- [10] 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.
- [11] 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.

- [12] 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.
- [13] 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
- [14] H. Bulut and S. Korukoglu, "Analysis and Clustering of Movie Genres", Journal Of Computing, Volume 3, Issue 10, pp.16 -23, October 2011
- [15] 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.
- [16] Large Movie Review Dataset, Acquired from Stanford AI Lab:(<http://ai.stanford.edu/~amaas/data/sentiment>).

