Genre-Based Movie Analysis from Viewer Reviews

Thilagavathy.N[1], Kannan.U[2], Gothandapani.M[3], Mathan.S[4]

[1] *Associate Professor, Sri Manakula Vinayagar Engineering College, Pondicherry-605104, India*
[2,3,4] *Dept. Of Information Technology, Sri Manakula Vinayagar Engineering College, Pondicherry-605104, India*
[1] thilagarajsanran@gmail.com
[2] kannan.u2791@gmail.com
[3] gothan05@gmail.com
[4] lsmathan@gmail.com

Abstract - **The Genre of the movie expresses the content of the movie. But in current system, the genre is expressed only in single aspect. This expresses the major content of the movie but not the entire content. So, the genre-based analysis is done to reveal the entire content of the movie in various genres. The analysis is done using the viewer reviews. First, a genre identification model is proposed to learn genre-related terms of each genre that are used for genre identification. Second, the polarity analysis is done to positive, negative and neutral words. In addition, the co-occurrence value of negative words along with the genre related terms are identified. Finally, the percentage of genres in the movie is calculated.**

I. INTRODUCTION

The movie reviews plays a very important role in the movie domain. There are different genres in the movie domain. The Viewer reviews contain much useful information about these genres. Websites like IMDB, BehindWoods offer the facilities for the viewers to provide the review for the movies. Genre-based analysis can be done for a movie from these viewer reviews. This is performed without asking any questions to the users. The reviews are provided for many online products. But plenty of reviews are provided in the movie domain only. So, the movie domain is selected. We analyze unlabeled free-form textual customer reviews and generate the genre based results. If the value of comedy genre is 0.25, then 25% of the particular film contains the comedy aspect. In this system, 3 modules are presented. First, Genre identification model is proposed to learn Genre-related terms of each genre that are used for genre identification. Second polarity analysis is done to identify the polarity of the genre related term. Finally the value is displayed as graph for easy understanding. The existing system can be easily implemented in any domain. The reviews are given for all the products available in online. Since the movie domain has great availability of such reviews, we opted to choose it. The proposed system deals with the English reviews. As stated earlier, the proposed system consists of 3 modules. First one is Genre identification model. Second, the polarity analysis. The third is graph representation. The earlier system is performed in the sentence level. Thus the aspect based sentence segmentation is needed. For this purpose, the grid search algorithm is used. In this system, we eliminated the aspect based sentence segmentation, since we consider the words rather the sentences.

II. GENRE IDENTIFICATION MODEL

In this section, 2 sub modules involved. First one is the Input processing module to prepare the input words. Second is the Bootstrapping Framework model to identify the genre of the word. The reviews collected manually from various movie websites are to be preprocessed for genre identification. The initial step is to identify the part of the speech of the term. Since the previous researches illustrates that (noun|verb|adverb|adjective) plays a major role in the reviews, the terms with these POS must be identified. This can be done by various available tools. We adopt the Stanford NLP (natural language processing) POS tagger for identifying the term of above specified POS.

The terms are extracted after POS tagging and are to be given as input to the next module. Before all these process, the initial seed set need to be developed manually. From available reviews from various movie websites, the genre related words are picked manually. The better initial seeds, the better the output results.

A. *Bootstapping:*

Before starting all process, the different genres of the movies are to be identified and to instruct the learning system, the initial seed set is created i.e. the words related to the genre are selected. Selecting an appropriate number of good seeds for bootstrapping is an important step, and how to select the best seeds for bootstrapping in real-world applications is still an open question.

In our solution, 500 frequently used nouns, verbs, and adjectives were first automatically extracted for manual selection of seeds. Therefore, to minimize human efforts, in each bootstrapping-based GRT learning algorithm, five seeds were manually chosen for each genre for they appear frequently in the unlabeled corpus. To evaluate the bootstrapping results, all learned GRTs for each aspect are ranked in descending order of their importance scores and the scores are calculated using *RlogF* method. The RlogF formula is given as follows,

$$\text{RlogF}(s) = \log \text{freq}(s, S) * \{\text{freq}(s, S)/\text{freq}(s)\}$$

where S is the current seed set, freq(s, S) is the frequency of co-occurrence of s and s within a given limited context (i.e., k words to left or right of s), freq(s) is the frequency of co-occurrence of s in the corpus. The general bootstrapping framework for genre-related term learning consists of the following steps, as shown in,

Algorithm 1:

Initial Seed:

A small number of manually chosen seed genre related terms, namely $G=\{g_1,g_2,\ldots,g_m\}$ for genre a and unlabeled data set X.

- Extraction of candidate GRT extraction: From the unlabeled data X, extract nouns, verbs, adjectives, adverbs to form candidate GRT set Ω
- Iterative Bootstrapping:

  1. Calculate the value score RlogF for each candidate GRT with respect to G.
  2. Select the candidate with the highest RlogF score to augment G, and remove it from set Ω.

  Until a desirable number of GRTs have been learned.

We score each candidate GRT s with the RlogF metric. The standard MAB algorithm favors the GRT with a high n(s) score and a low φ(s) value for learning, in which the importance score assigned to an GRT s for genre a_i can be calculated by

$$score_i(s) = n_i(s) * (1 - \varphi(s))$$

where $n_i(s) = 1 - (r_i(s) / |G_i|)$

$G = \{g_{i1}, g_{i2}, \ldots, g_{im}\}$ is the GRT set of genre a_i learned. Notice that gij is learned in the jth iteration, |Gi| indicates the number of GRTs in G_i, and $r_i(s)$ represents the rank of s in G_i, indicating in which iteration it was learned.

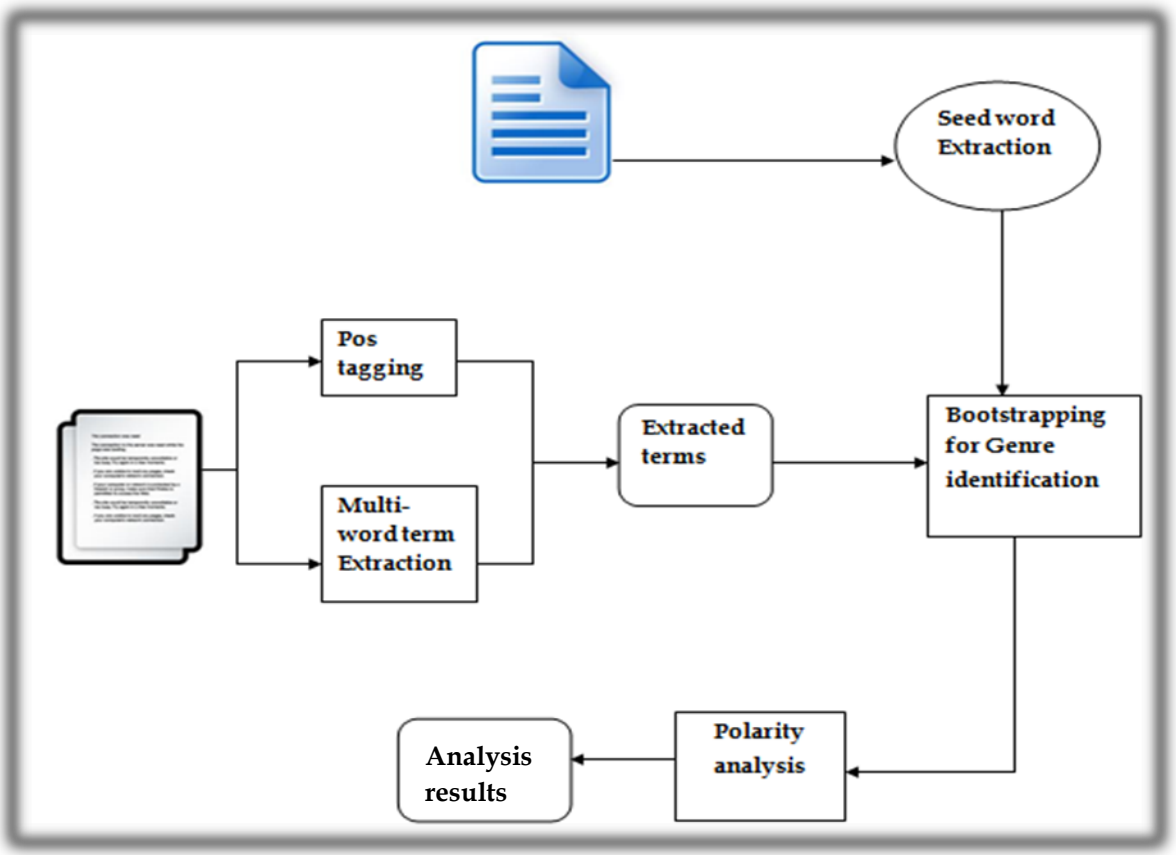
B. Polarity analysis

Here the polarity of the term is identified by using the polarity analysis tool. The co-occurrence of the GRT term with the negative prefix is calculated and the score is subtracted with the actual score of the GRT term. The polarity analysis tool analyses the positive, negative and neutral percentage of the term. These positive values are alone concentrated and multiplied individually the scores and overall the value of the individual genre is calculated.

C. Graph representation:

The calculated value is represented in the form of a graph. The current system is given as in the form of star rating. The alternate solution will be expressing the entire content of the movie in the bar diagram.

III. ARCHITECTURAL MODEL



IV. FUTURE WORK

The previous work in the review field is done with the Chinese language, which is entirely different from the other languages. In our proposed system, we carried out with the English language and the domain we selected is the movie domain. Since there are movies in various regional languages, there will be reviews available in the respective regional languages. So we will consider including to develop a system that will consider the available reviews in all regional languages of a particular movie in various genres.

V. CONCLUSION

The proposed system will enable the user to identify the content of the movie in the various genres. We removed the sentence segmentation method, since our proposed system works in the word level. This will improve the performance and the accuracy of the results.

References:

[1] F.Y.Y. Choi, "Advances in Domain Independent Linear Text Segmentation," Proc. First Meeting North Am. Chapter Assoc. for Computational Linguistics, pp. 26-33, 2000.

[2] K.W. Church, "Char Align: A Program for Aligning Parallel Texts at the Character Level," Proc. 31st Ann. Meeting Assoc. for Computational Linguistics, pp. 1-8, 1993.

[11] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature Subsumption for Opinion Analysis," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 440-448, 2006.

[12] E. Riloff, J. Wiebe, and T. Wilson, "Learning Subjective Nouns Using Extraction Pattern Bootstrapping," Proc. Seventh Conf. Natural Language Learning at HLT-NAACL, 2003.

[3] P. Fragkou, V. Petridis, and A. Kehagias, "A Dynamic Programming Algorithm for Liner Text Segmentation," J. Intelligent Information System, vol. 23, no. 2, pp. 179-197, 2004.

[4] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic Recognition of Multi-Word Terms: The C-Value/NC-Value Method," Int'l J. Digital Libraries, vol. 3, pp. 115-130, 2000.

[5] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining Customer Opinions from Free Text," Proc. Sixth Int'l Symp. Intelligent Data Analysis, pp. 121-132, 2005.

[6] V. Hatzivassiloglou and K. McKeown, "Predicting the Semantic Orientation of Adjectives," Proc. 35th Ann. Meeting of the Assoc. for Computational Linguistics and Eighth Conf. European Chapter of the Assoc. for Computational Linguistics, 1997.

[7] E. Milios, Y. Zhang, B. He, and L. Dong, "Automatic Term Extraction and Document Similarity in Special Text Corpora," Proc. Sixth Conf. Pacific Assoc. for Computational Linguistics, pp. 275- 284, 2003.

[8] M. Krauthammer and G. Nenadic, "Term Identification in the Biomedical Literature," J. Biomedical Informatics, vol. 37, no. 6, pp. 512-526, 2004.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," Proc. ACL- 02 Conf. Empirical Methods in Natural Language Processing, 2002

[10] J.C. Reynar, "An Automatic Method of Finding Topic Boundaries," Proc. 32nd Ann. Meeting Assoc. for Computational Linguistics, pp. 331-333, 1994.

[13] I. Titov and R. McDonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization," Proc. Assoc. for Computational Linguistics, pp. 308-316, 2008.

[14] X. Wan, "Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 553-561, 2008.

[15] B. Wang and H. Wang, "Bootstrapping Both Product Properties and Opinion Words from Chinese Reviews with Cross-Training," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 259-262, 2007.

[16] T. Zagibalov and J. Carroll, "Unsupervised Classification of Sentiment and Objectivity in Chinese Text," Proc. Third Int'l Joint Conf. Natural Language Processing, pp. 304-311, 2008.

