Sentiment Analysis of Movie Reviews Using Heterogeneous Features

Rachana Bandana
Department of Computer Engineering,
Dharmsinh Desai University,
Nadiad, India.
rbaldania2103@gmail.com

Abstract—**Human disposition has always influenced by others suggestion and reviews. People are always eager to know other's reviews for their profit but, every website contains a very large amount of review text, the average human reader will have trouble in identifying relevant sites, extracting and abstracting the reviews so they cannot reach to the right decision in less time that is why automated sentiment analysis systems are required. In the proposed approach, heterogeneous features such as machine learning based and Lexicon based features and supervised learning algorithms like Naive Bayes (NB) and Linear Support Vector Machine (LSVM) used to build the system model. From implementation and observation, conclude that using proposed heterogeneous features and hybrid approach can get an accurate sentiment analysis system compared to other baseline system. In future for big data, we can use these heterogeneous features for bulding advance and more accurate models using Deep Learning (DL) algorithms.**

Keywords—Natural Language Processing (NLP); Text Mining; Information Retrieval; Data Mining; Big Data; Sentiment Analysis; Opinion Mining; Machine Learning (ML); Deep Learning (DL); SentiWordNet (SWN); Lexicon; Hybrid; Python; Natural Language Processing Toolkit (NLTK).

I. Introduction

Since early 2000, sentiment analysis is the most vigorous research areas of Data Mining and Natural Language Processing (NLP). Human disposition has always influenced by others suggestion and reviews [1]. That is why our reviews are very much influenced by other's reviews, and whenever we need to make a decision, we often seek out other's reviews. When we need to check reviews, we will started trying to find reviews in many digital platforms such as social media , reviews site, forum discussions, blogs, and micro blogs, though every website contains a very large amount of review text, the average human reader will have trouble in identifying relevant sites, extracting and abstracting the reviews so we cannot reach to the right decision. This is not only true for any individual business person but also for organizations, companies, political parties that is why people need automated smart sentiment analysis system which can accurately give correct sentiment and relevant information in less time for their benefits.

Definition of Sentiment is an attitude, emotion or mood, etc., and Definition of Sentiment analysis is to identify the polarity of review text or the subjective and the emotions of a particular topic in document or sentence [2].

Mine people's review and feelings toward any subject matter of interest, which is the task of sentiment analysis [3].

Now a day, sentiment analysis can apply to almost all possible domains like products, services for social events and political elections, market research, social media, advertising, recommendation systems, email filtering, stock market prediction, upcoming movie reviews sentiment prediction, book reviews sentiment, etc.

II. Proposed Approach

In the proposed system identified sentiment orientation from review text documents using a hybrid approach. The hybrid approach means a combination of Machine learning and Lexicon-based (knowledge-based) approach.

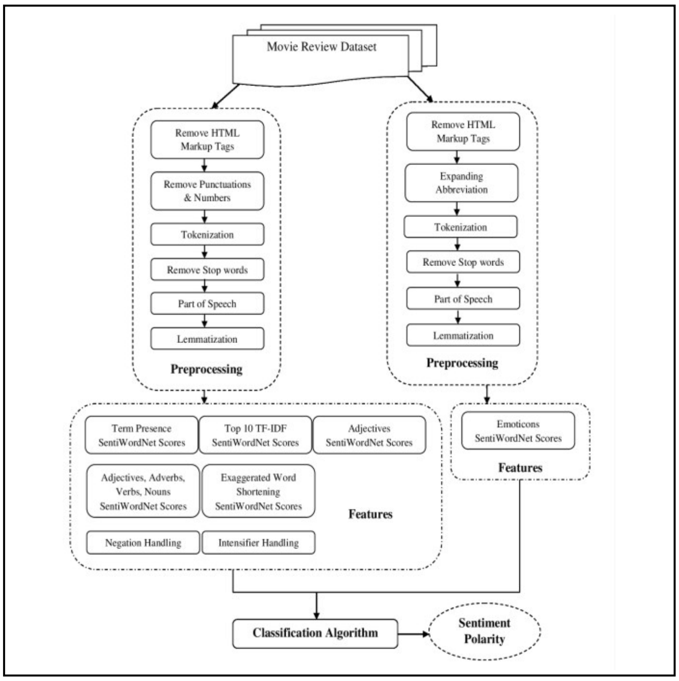

Fig. 1. Proposed Approach System Block Diagram.

978-1-5386-5550-4/18/$31.00 ©2018 IEEE

As we can see in Fig. 1, Block diagram of proposed approach has major five components, which are movie review data set, preprocessing, features selection and extraction, classification algorithms and sentiment polarity. As an input, manually created movie review text documents are used, but when we collect the data from the web there are many irrelevant and unprocessed data so they need to preprocess using different preprocessing techniques than feature selection and extraction will be done and After getting features matrix, we have to applied this matrix to different supervised learning classifiers such as, Naive Bayes and Linear Support Vector Machines and from that predict the sentiment label which gives a review text polarity orientation in a positive or negative manner. Following are the major components of the proposed approach:

A. Movie Review Data set

The Dataset is manually created and labeled using different sources like BookMyShow, IMDB, Rotten Tomatoes, Netflix, etc. and it contains text movie reviews which comprised of Tab-separated values (TSV) files in which, there is an id column header is unique for identifying document uniquely and sentiment header is expected output for reviews which is two class labels 1 and 0, where 1 means positive class and 0 means negative class and review header is contains the movie reviews. This TSV file will different for training and testing, in the testing TSV file, there will no sentiment label because it will predict by algorithm. Basically two movie reviews data set created, in first 250 movie review text documents where, 250 for training and 100 for testing and validation. In second 300 movie review text documents where, 300 for training & 150 for testing and validation used. Each sentence contains around 27 to 103 words.

B. Preprocessing

Data collected from the web are not cleaned and therefore the data set consists of unnecessary and redundant information so that we have to remove unnecessary information for improving accuracy and efficiency for system that is why different preprocessing techniques like Remove HTML Markup Tags, Remove Punctuations & Numbers, Expanding Abbreviation, Tokenization, Remove Stop words, Part of Speech, Lemmatization used but sometimes for future requirements, we have to apply different preprocessing step like as we can see in Fig. 1. Proposed system block diagram there are two preprocessing approach, in the first case, remove punctuation and didn't use expand abbreviation feature and the second case is specially for creating emoticons feature because in emoticons (ex., :) or :/) because in second case, we have to use punctuations and special symbols, so in this case we are not going to remove punctuations and special symbols and we also used expanded abbreviation feature (ex. don't will be replaced by do not), but this is an optional feature.

C. Features

A Feature is used mainly for reduce vocabulary size, noise features and increased classification accuracy. Heterogeneous features created using a combination of lexicon like SentiWordNet, WordNet, etc., and machine learning like bag of words, TF-IDF, etc.

In this Research work, we have created a special featue using SentiWordNet, ‘SentiWordNet Average on Review’: for each review document, SentiWordNet scores determined by calculating the average of the total of positive and negative SWN scores of all words in a document, so using this technique will be useful in creating many features. Here are the heterogeneous features:

1) SentiWordNet Scores: SentiWordNet Average on Review technique used (as we discussed in section II), used to identify score for only adjectives, adverbs, verbs and noun words rather than all words in document because all these are important indicators of sentiment in review text.

2) SentiWordNet Scores (Adjectives only): SentiWordNet Average on Review technique used to identify score for only adjectives words rather than all words in document because all these are important indicators of sentiment in review text.

3) Term Presence SentiWordNet scores: Binary value indicates a vocabulary word present or not in the document, when the word occurs in the document, it is one and zeros for all words that appear in the document, calculate score of SentiWordNet Average on Review technique where word present but, here also considered only adjectives, adverbs, verbs and nouns.

4) Top 10 TF-IDF SentiWordNet: From TF-IDF top 10 high informative words which should be adjectives, adverbs, verbs and nouns selected and also compute the score of SentiWordNet Average on Review, but here also considered only adjectives, adverbs, verbs and nouns.

5) Exaggerated word shortening: Usually people use repeated letters in words like 'wowwwww' to show their intensity of sentiment. But, these words are not present in the SentiWordNet hence the extra letters in the word must be eliminated. So after that change, calculated SentiWordNet Average on Review, but here also considered only adjectives, adverbs, verbs and nouns.

6) Negation Handling: Sentence may contain negative modifiers like 'not' 'never', 'no' which change the meaning of that review so that should not happen that is why handled negation.

7) Intensifier handling: People usually use intensifiers in reviews to express their sentiments deeply. The presence of the words like 'very', 'really' and 'extremely' in negative and positive reviews make the adjective and adverb stronger that is why to find stronger sentiment we should handle intensifier.

8) Emoticons only: Many blogs and social posts make use of emoticons in order to convey sentiment, making them very useful for sentiment analysis. A range of about 10 emoticons, including :) :(|-{ :D :-| :P :* :-} etc. are replaced with either a HAPPY or SAD keyword and calculated SentiWordNet Average on Review, but here also considered only adjectives, adverbs, verbs and nouns. Please note that for implementing Emoticons features don't remove punctuation in preprocessing technique.

D. Classification Algorithms

Sentiment analysis involves classifying reviews in text into categories based on sentiment polarity. Polarity in the review text means the state of having two opposite reviews. This polarity can be positive or negative. Here, in proposed approach document-level sentiment analysis classification used and here applied heterogeneous features to supervised machine learning algorithms like Naive Bayes (NB) and Linear Support Vector Machine (LSVM) to learn and classify text reviews into a positive and a negative category. Here for this system choosed a supervised learning algorithm because we can easily get enough dataset for training.

E. Sentiment Polarity

From supervised classifier, we can predict the sentiment class label, it will be positive or negative class means given review text document will positive or negative.

III. Implementation and observations

For the implementation of proposed systems, Python 3.4 and Natural Language Processing Toolkit (NLTK) are used. NLTK, is a set of functions and programs for statistical natural language processing (NLP) for the Python programming language. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems [22].

TABLE I. Observations

Dataset	Algorithm	Accuracy	Precision	Recall
250 Training and 100 Testing	Naive Bayes	89%	89%	88%
250 Training and 100 Testing	Linear SVM	76%	82%	76%
300 Training and 150 Testing	Naive Bayes	84%	83%	84%
300 Training and 150 Testing	Linear SVM	79%	85%	80%

Here used different large and small data sets for different preprocessing techniques, heterogeneous features and classification algorithms and observe various results which are, shown in TABLE I.

IV. Conclusion and Future Work

From experiments, can conclude that using these heterogeneous features we can get better results rather than using only machine learning or lexicon based features and also using the Naive Bayes algorithm achieved tremendous accuracy even in a small amount of training data as especially compared with Linear SVM for these heterogeneous features.

In the future, we can experiment with different preprocessing techniques, heterogeneous features, supervised and unsupervised algorithms for developing a more accurate system. For handling large data; we can use proposed heterogeneous features and deep learning features such as Word2Vec, Doc2Paragraph and Word Embedding apply to deep learning algorithms such as Recursive Neural Network (RNN), Recurrent neural networks (RNNs) and Convolutional deep neural networks (CNNs) to get remarkable result.

Acknowledgment

I will take this opportunity to thank Dean, Head of Department (HOD) and my guide Dr. Brijesh Bhatt of my institute for providing me the guidance, environment and support to complete this research work.

References

[1] Rachana. Baldania, "Sentiment analysis approaches for movie reviews forecasting: A survey," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-6.

[2] Singh, Rajni, and Rajdeep Kaur. "Sentiment Analysis on Social Media and Online Review." International Journal of Computer Applications 121, no. 20 (2015).

[3] Liu, Bing. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015.

[4] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86. Association for Computational Linguistics, 2002.

[5] Mudinas, Andrius, Dell Zhang, and Mark Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, p. 5. ACM, 2012.

[6] Khan, Aamera ZH, Mohammad Atique, and V. M. Thakare. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE) (2015): 89.

[7] Hamouda, Alaa, and Mohamed Rohaim. "Reviews classification using sentiwordnet lexicon." In World Congress on Computer Science and Information Technology. 2011.

[8] Shukla, Ashish. "Sentiment Classification And Analysis Using Modified K-Means And Naive Bayes Algorithm." PhD diss., Uttar Pradesh Technical University, 2015.

[9] Guerini, Marco, Lorenzo Gatti, and Marco Turchi. "Sentiment analysis: How to derive prior polarities from SentiWordNet." arXiv preprint arXiv:1309.5843 (2013).

[10] Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." In 9th. IT & T Conference, p. 13. 2009.

[11] Yadav, Shailesh Kumar. "Sentiment analysis and classification: A survey." International Journal of Advance Research in Computer Science and Management Studies 3, no. 3 (2015).

[12] Annett, Michelle, and Grzegorz Kondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." In Advances in artificial intelligence, pp. 25-35. Springer Berlin Heidelberg, 2008.

[13] Bhoir, Purtata, and Shilpa Kolte. "Sentiment analysis of movie reviews using lexicon approach." In 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-6. IEEE, 2015.

[14] Ouyang, Xi, Pan Zhou, Cheng Hua Li, and Lijun Liu. "Sentiment Analysis Using Convolutional Neural Network." In Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on, pp. 2359-2364. IEEE, 2015.

[15] Singh, V. K., R. Piryani, Ahsan Uddin, and P. Waila. "Sentiment analysis of Movie reviews and Blog posts." In Advance Computing Conference (IACC), 2013 IEEE 3rd International, pp. 893-898. IEEE, 2013.

[16] Vijayarani, S., Ms J. Ilamathi, and Ms Nithya. "Preprocessing Techniques for Text Mining-An Overview." vol 5: 7-16.

[17] Katariya, Ms Nikita P., M. S. Chaudhari, B. Subhani, G. Laxminarayana, Kalyani Matey, Ms Archana Nikose, Sonali A. Tinkhede, and S. P. Deshpande. "Text Preprocessing For Text Mining Using Side Information." (2015).

[18] Muhammad, Ajmal, Nirmalie Wiratunga, and Robert Lothian. "A Hybrid Sentiment Lexicon for Social Media Mining." In Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on, pp. 461-468. IEEE, 2014.

[19] Pouransari, Hadi, and Saman Ghili. "Deep learning for sentiment analysis of movie reviews." (2014).

[20] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In LREC, vol. 10, pp. 2200-2204. 2010.

[21] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2, no. 1-2 (2008): 1-135.

[22] Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, p. 1642. 2013.

[23] https://en.wikipedia.org/wiki/Natural_Language_Toolkit1989.