A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts

Authors: Bo Pang and Lillian Lee
Year: 2004
Journal: Proceedings of the 42nd Association for Computational Linguistics.

Report:

1. **Abstract:**

"Sentiment analysis seeks to identify the viewpoint(s) underlying a text span; an example application is classifying a movie review as "thumbs up" or "thumbs down". To determine this sentiment polarity, we propose a novel machine-learning method that applies text-categorization techniques to just the subjective portions of the document. Extracting these portions can be implemented using efficient techniques for finding minimum cuts in graphs; this greatly facilitates incorporation of cross-sentence contextual constraints."

Introduction

- The work primarily focuses on the sentiment polarity extraction by applying text categorization approaches only to the subjective parts in the document.
- Despite of facts, the authors want to create the summarizations on the subjective opinions which would be very well useful in the corporate organizations, product reviews etc.,
- Also, they have explored the minimum cut algorithm to bring up the efficient means of integrate the contextual information.

Previous works

- The need of neutral reviews had been a need for many of the products, films, etc., as the previous papers mostly dealt with the subjective reviews of the documents, this paper created an algorithm to result the best review out of subjectivity.
- Even the earlier works had mostly focused on indicative meanings of the content provided.
- Moreover the authors had used the graph algorithm called as minimum cuts, which resulted in most efficient and intuitive results.

Content

- Generally, the content is used to derive the sentiment analysis rather than the topic based structures. For this reason the SVMs are adapted to extract the sentiment as done in the previous works.
- Initially a 'n' number of sentences were considered and they were checked based on the subjectivity wise.

- This resulted in number of subjectivity related sentences and those sentences are extracted. Later a Machine learning technique is used whether the subjectivity sentences lead to positive or negative polarity.
- The use of minimum cut leads to 'n' terms of x are taken as $x_1,x_2,x_3,.....x_n$ and they are classified into two classes named C_1 and C_2.
- This method is adapted to minimize the cost of the process, which resulted in optimizing problem to find the minimum cuts.
- After classifying into two classes they represented scores for every element as individual scores and association scores.
- Individual scores define the score of each element to relate to some class, where as the associative scores define the scores of the element to be in one class based on the other element.
- The graphs have been created based on the score achieved independently based on item specific and pairwise scores.
- **Subjectivity dataset** is the dataset of labeled sentences to train the subjectivity detectors. They have extracted the data from the web of around 5000 movie review snippets.
- **Subjectivity detectors** are the default polarity classifiers which are taken from the subjectivity dataset as described before. They even constructed the cut-based subjectivity dataset.
- The graph is drawn to denote the between the item wise and pairwise relationships of each sentence of the dataset.
- Most of the last sentences were considered and found the relationships of all the sentences, which would normally contains the maximum summary of the content.

Results

- A satisfying improvement had been shown in usage of subjectivity extraction by polarity classification.A good accuracy of 86.4% from 82.8% had been achieved by taking Naive Bayes as the subject detector along with the Naive Bayes document polarity classifier.
- The reduction of the subjectivity of the original documents is also conducted and gained very good texts' sentiment accuracy.
- The approach of taking the last sentences also resulted in the very good accuracy compared to the initial sentences of the document, which is because most of the summarized data will be available at the end of the document.
- It is said that the addition of the proximity information to the normal Naive Bayes extraction and SVM extraction also resulted in the higher accuracy comparatively.
- Although, better results occurred with the proximity information, without that the Naive Bayes performed well than the SVM extraction.
- Finally, the actual motivation of the paper is achieved in creating the very compact summaries of the actual reports with as much less of subjectivity oriented.

Discussion and thoughts

- The paper is is an extension for the betterment of sentimental education which was tried by the new approach called minimum cut algorithm.

- They have seen the better results from the sentences of end of the document, which may take many further calculations to reach there and had to all the values before.
- For that reason, they stated that it would have been better if the authors start writing the summaries at the beginning, which would result in turbulence. Rather it would been great if they would have started verifying the sentences from the bottom.

Conclusion

- On the whole, the paper had not been a pioneer of an algorithm or the process, they indeed used the existing procedures to increase the efficiency.
- The paper also resulted in a significant increase of accuracy with the new algorithm, which had been one of the motivation.
- The paper would been a good for the people who can continue the works in the opinion extraction and the semantic orientation based on the polarity classification.
- The neutralization of reviews which are very essential rather than subjective is also achieved by resulting in the compact reviews.
    -