NOVEL ALGORITHM FOR SPAM DETECTION AND SENTIMENT AND FEATURE LEVEL ANALYSIS OF REVIEWS

Rishabh Chand Sharma¹, Manish Mathuria²
¹M.Tech. Scholar, ²Head of CS Dept, Maharishi Arvind College of Engineering & Research Center

Abstract: The Review submission is the normal process when we talk about the social media like facebook, google+ etc. These reviews are very useful for companies for improving the products specifications and quality. But always the reviews are not of use as the internet is also in attack by spammers and people writing spam reviews. So the main problem is to filter these reviews. In our dissertation, we have proposed the novel algorithm for detection of the spam reviews and together with the detection of the spam reviews, we have extended our work in classification of the reviews like positive reviews and negative reviews. In our work we have used spam and ham dataset for the identification of the reviews as spam or ham. And after the normal review is detected then using the sentiment analysis we have classified the reviews.

1. INTRODUCTION

It has become a common practice for people to read online opinions/reviews for different purposes. For example, if one wants to buy a product, one typically goes to a review site (e.g., amazon.com) to read some reviews of the product. If most reviews are negative, one will almost certainly not buy it. Positive opinions can result in significant financial gains and/or fame for businesses, organizations and individuals. This, unfortunately, gives strong incentives for opinion spamming.

1.1 Opinion Spamming

It refers to "illegal" activities (e.g., writing fake reviews, also called shilling) that try to mislead readers or automated opinion mining and sentiment analysis systems by giving undeserving positive opinions to some target entities in order to promote the entities and/or by giving false negative opinions to some other entities in order to damage their reputations. We believe that as opinions on the Web are increasingly used in practice by consumers, organizations, and businesses for their decision making, opinion spamming will get worse and also more sophisticated.

1.2. Feature Level Analysis:

Both the document level and the sentence level analyses don’t discover what exactly people liked and did not like. Aspect level performs finer-grained investigation. Aspect level was earlier called feature level (feature-based conclusion mining and summarization). Instead of taking a gander at language builds (documents, sections, sentences, clauses or phrases), aspect level directly takes a gander at the assessment itself. It is based on the idea that an assessment consists of a sentiment (positive or negative) and a target (of conclusion).

1.3. Sentiment Classification Techniques

In general, Sentiment Classification should be possible with three techniques machine learning (ML) approach, lexicon based approach and hybrid approach.
related text fragment to the user. The authors discussed about some existed research work as many search engine retrieved facts through keyword matching, popularity etc.

III. PROBLEM DESCRIPTION

This chapter incorporates the issue definition and the destinations alongside the subtle elements of the procedures utilized as a part of the proposed work.

Objectives

- Spam Filter of the reviews
- To gather the reviews for motion picture space from various social locales.
- To perform information-preparing as
  - Tokenization,
  - Stop word evacuating
  - Stemming and grammatical feature tagging on gathered reviews for information readiness.
- To concentrate every one of the components from the reviews and store in the database.
- To decide the polarity of the basic sentence and compound sentences at highlight level utilizing proposed calculation.

To look at our proposed approach utilizing existing sentiment examination device (Opinion Finder, SentiWordNet, and WordNet spread).

![WordNet](image2)

IV. PROPOSED METHODOLOGY

4.1 Algorithm Adopted For Spam Detection

Step 1: Read the document file containing the Review

Step 2: Read the dataset containing the HAM keywords.

Step 3: Read the dataset containing the SPAM keywords.

Step 4: Analyze the document for the SPAM and HAM keywords on the basis of the occurrences.

4.2 Algorithm Adopted For Classification

Proposed model for Movie Review Analysis using the Sentiments scoring based on SentiWordNet and feature level analysis of the documents as shown in figure 4.1

Step 1: Input

Input Review document for generating review result that whether the review is positive or Negative (.txt file).

Step 2: Segmentation

Divide the document into sentences using segmentation.

Step 3: Feature Level Filtering

We will read the file is the feature is all then all the files will consider for the review analysis otherwise the filtration is done. And the algorithm is shown in Algorithm-1 [22].

ISFEATURE (line, feature)

[Perform feature level extraction]
1. Check for the Feature and its Synonyms.
2. If the line contains words which are feature itself or its synonyms
3. Return the line granted for review analysis.

[End of for loop]

Algorithm-1: Feature Level Filtering Algorithm

Step 4: Tokenization

Each sentence is divided into tokens.

An example: Friends, has been colouring and roman lends me, your field; Hence after tokenization we get: Friends has been colouring and roman lends me your field. Basically we need to omit the commas, punctuations, (carefully apostrophes), question marks etc [23].

Step 5: POS tagging

The pos tagging is the tagger which specify the token as nouns, verbs, adverbs, adjectives [24] show in Algorithm-2.

STEMTAGG (line)

[Perform stemming and then tagging]
1. Using the WORDNET API, the base for of words is obtained
2. Tagged the line using the MAXENTTAGGER class.
3. Return the finally tagged line.

Algorithm-2: Stemming & Tagging Algorithm

Step 5: Splitting the Line and processing for line score.

In this we will split the line into the array and we will find the score of the each word in order to get the score for the complete sentence. And this score will further led to the scoring for the entire document. The scoring is done using the SentiWordNet and here we will use the SentiWordNet library for getting the scoring as described in the Algorithm-3.

SENTISCORE (Word, POS_tagg)

[Perform scoring based on SentiWordNet]
1. Call the sdata.txt library file for SentiWordNet scoring.
2. Call extract (word, tag) for getting the score for the word.
3. Return the score.

Algorithm-3: Scoring Word using SentiWordNet

Step 6: Intensifier Handling

In this we will examine that the word which we are reading from the line is intensifier or not, if it is intensifier then the score is to be handled accordingly, as the intensifier will further enhance the score show in Algorithm-4.

CHECKINTENSIFIER (Word)

[This algorithm will check whether the word is the intensifier or not.]
1) Read the File “intensifier.txt” into ifstream
2) Repeat till EOF (End of File)
3) Read INTENSIFIER
4) If Word is INTENSIFIER then :
   Return true
Else
   Return false
[End of If structure]
[End of inner for loop]
VI. TEST RESULT

6.1 Comparison Test Results for Negation Handling

<table>
<thead>
<tr>
<th>Review Text</th>
<th>Proposed Work</th>
<th>Base paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>There was no good story in that movie.</td>
<td>-0.659</td>
<td>0.608</td>
</tr>
<tr>
<td>There was good story in that movie.</td>
<td>0.608</td>
<td>0.608</td>
</tr>
<tr>
<td>The story of the movie was not good.</td>
<td>-0.684</td>
<td>-0.042</td>
</tr>
<tr>
<td>The story of the movie was good.</td>
<td>0.582</td>
<td>0.582</td>
</tr>
<tr>
<td>The music of the songs was not bad.</td>
<td>0.953</td>
<td>-0.812</td>
</tr>
<tr>
<td>The music of the songs was bad.</td>
<td>-0.187</td>
<td>-0.187</td>
</tr>
<tr>
<td>The hero acting was not bad in Bajirao Mastani movie.</td>
<td>0.712</td>
<td>-1.053</td>
</tr>
<tr>
<td>The hero acting was bad in Bajirao Mastani movie.</td>
<td>-0.428</td>
<td>-0.428</td>
</tr>
</tbody>
</table>

Table 6.1 Comparison Test Results for Negation Handling

6.2. Comparison Test Results for Intensifier Handling

<table>
<thead>
<tr>
<th>Review Text</th>
<th>Proposed Work</th>
<th>Base paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>The movie was very good.</td>
<td>0.717</td>
<td>0.612</td>
</tr>
<tr>
<td>The movie was good.</td>
<td>0.6337</td>
<td>0.6337</td>
</tr>
<tr>
<td>The songs were too bad.</td>
<td>-0.487</td>
<td>-0.653</td>
</tr>
<tr>
<td>The songs were bad.</td>
<td>-0.570</td>
<td>-0.570</td>
</tr>
<tr>
<td>The movie was directed very badly.</td>
<td>-0.349</td>
<td>-0.234</td>
</tr>
<tr>
<td>The movie was directed badly.</td>
<td>-0.432</td>
<td>-0.432</td>
</tr>
</tbody>
</table>

Table 6.2 Comparison Test Results for Intensifier Handling

The percentage is calculated on the average of the number of reviews compared in both the implementation and some samples are shown in the tables presented above.

REFERENCES

[12] Dr. Muhammad Shahbaz1, Dr. Aziz Guergachi2, Rana Tanzeel ur Rehman3,” Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text” 978-1-4799-3010-9/14/$31.00 ©2014 IEEE
[14] J. Ashok Kumar, S. Abirami, and AN


[17] Mohsen Farhadloo , Erik Rolland,” Multi-Class Sentiment Analysis with Clustering and Score Representation” 2013 IEEE 13th International Conference on Data Mining Workshops DOI 10.1109/ICDMW.2013.63


[19] Ms.K.Mouthami, Ms.K.Nirmala Devi, Dr.V.Murali Bhaskaran “Sentiment Analysis and Classification Based On Textual Reviews”


[22] Na Fan, Wandong Cai, Yu Zhao ”Research on the Model of Multiple Levels for Determining Sentiment of Text” IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application,2008.


