Information Retrieval (CS F469) Assignment 4

2015A7PS0133H Mukund Kothari
2015A7PS0136H Aayush Barthwal
2015B4PS0546H Rohitt Vashishtha
2015A7PS0047H Tushar Aggarwal

A Report on Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis

- **Problem Statement**:
  The number of clients of on-line shopping sites is persistently expanding. Such sites often provide the clients the interface to give comments and appraisals to the items being sold on these sites. This data can be valuable as suggestion for different clients in settling on their buy choice**. This paper explores the issue of anticipating rating in view of clients' remarks.**

  Because of the continues growth of on-line shopping sites, many Internet users think that it's extremely hard to settle on choice on their shopping needs. The user needs to assess numerous comparative items with various highlights, quality and costs before settling on a buy choice. Dissimilar to offline shopping, where the clients can evaluate the items physically, these users are dependent on the special pictures or recordings to settle on the product that they would like to purchase.

  All the more as of late, on-line shopping sites, for example, Flipkart.com, enable the Internet clients to give rating for items that are being sold on the site. The rating will show the level of fulfillment of the clients for the specific items. Furthermore, those sites additionally give facilities to the clients to post comments with respect to the items. Such remarks can represent their opinion on various aspects of these items. Luckily, these comments and remarks can be utilized by the Internet users as an extra judgement criterion to help them in making the right decisions while avoiding counterfeit and deceptive products and sellers.

  By accepting that the appraisals and remarks are derived from the experience of other users' view on the same products, different users can settle on their choice on how much that particular product will meet his/her desire in light of other's involvement. Be that as it may, as the quantity of appraisals and remarks for an item expands, it becomes increasingly difficult for the users to investigate the volume of the data. A summarization device is thus valuable to help clients in breaking down the data before making a decision. Then again, the evaluations and the remarks

can likewise be utilized as another technique to send criticism to the organizations with respect to their products.

- **Solution Approach**:

A classifier based on Information Retrieval model is proposed for the prediction of ratings for any particular product. In addition, the effect of integrating sentiment analysis for the rating prediction is also investigated. Based on the results, an improvement in prediction performance can be expected with the integration of sentiment analysis into the model where an increase of around 54% is achieved.

A summary of ratings is very insignificant, for example, a normal rating, can be very informative for the clients to survey the general satisfactions on the item. Rating is regularly between a range of numbers, from 1 to 5. Thus, users can pick an item with higher normal rating for making on a buy choice. Then again, a summary of remarks is also troublesome as the users need to investigate the content from the remarks to find the general conclusion on the item. Sentiment analysis algorithm can be connected to the remarks to extricate the extremity of the content, regardless of whether are positive or negative. Once the extremity of the writings can be found, the remarks can be condensed in light of its extremity, for example, the level of the positive or negative. Thus, a summary of appraisals and remarks can lessen the complexity of data analysis for the users in making a purchase decision.

Numerous information retrieval (IR) models, for example, the Vector Space Model (VSM), are utilized to scan for relevant documents in response to a query by computing similarity between the text query and the text in documents. Such a model represents the documents and the query as vectors and the similarity is evaluated in view of the separation between two vectors. In this specific situation, it is easy to view that the comments with similar ratings are more like each other when contrasted with the comments with various ratings. All things considered, an IR model can be utilized to foresee the rating of item in light of the remarks from the users. Besides, it is accepted that the prediction model can be enhanced by the implementation of sentiment analysis to detect the extremity (negative or positive) of the comments.

In this paper, each rating is thought to be a class and there are 5 classes (rating 1 to 5). A regulated classifier based VSM is proposed to characterize content from the remarks into its rating class. The classifier is then extended to use sentiment analysis to improve the classification performance.

In order to utilize VSM as a text classifier, a document surrogate to represent each evaluating class should be developed. A document surrogate represents the text qualities of all comments in each evaluating class. At that point, the similarity score is computed between the new

comment to be classified and all five document surrogate for the rating classes. The class for the new comment is determined based on the highest similarity score.

Text Classifier based on VSM

The following steps are followed to construct a text classifier based VSM

- The document surrogates are constructed by aggregating all the comments for each class, such that:

  collection = $[V_1, V_2, ..., V_5]$; where V_i is the document surrogate for rating i.
- Tokenization of document surrogates is done and all the stop-words are removed followed by applying Stemming algorithms and finally indexing.
- Indexing of the document surrogates is done according to VSM Model. A vector of index terms is created for each document surrogate where the weight of each term in the vector is determined by the term frequency tf than the inverse document frequency idf.

  Here,

  $$\text{weighted term frequency } w_{tf} = \begin{cases} 1 + \log_{tf} & ; \text{if } tf >= 0 \\ 0 & ; \text{otherwise} \end{cases}$$

  Where tf is the frequency of the term in a document surrogate

  and,

  $$idf = \log(1+ (N/n_i))$$

  where idf is the inverse document frequency, N is the total number of document surrogates in the collection (i.e. 5 in this paper), and n_i is the number of document surrogates where the term i appears
- Finally, the weight of index terms in each vector is calculated based on the following equation:

  $$tf.idf = w_{tf} \times idf$$
- In order to determine the rating class for a new comment, a query is constructed based on the text from the new comment. A query vector is produced that consists of the weighted index terms. Similarity scores are calculated between the query and the document surrogates in the index by using cosine similarity score.

Extended text classifier based on VSM and sentiment analysis

It is easy to view that the comments with similar ratings are more like each other when contrasted with the comments with various ratings. Consequently, the comments are isolated into ten classes in view of the rating and the extremity (positive or negative) rather than five classes in the past segment. At that point, a sum of ten document surrogates is built by collecting the comments from each class.

$$\text{Collection} = \begin{bmatrix} (V_1^+, V_1^-) \\ \dots \\ (V_5^+, V_5^-) \end{bmatrix}$$

The classifier is built based on the steps from the previous section. There are two classes for each rating.

Sentiment Analysis

The model suggested above requires a sentiment analysis model to classify the comments as being either positive or negative. The following steps are taken to classify the comments:

- Tokenize comments into bag of terms
- The number of occurrence of the positive and negative terms in the comment is calculated based on the lexical dictionary
- The Polarity of a comment is determined based on the highest number of occurrence of the positive or negative terms in the comment.

The performance of the sentiment analysis is measured by positive predictive value, ppv, negative predictive value npv, and the classification accuracy.

Where

Ppv = (tp) / (tn+fp)

Npv = (tn) / (tn+fn)

and Accuracy = (tp+fn) / (tp+ tn + fp + fn)

where tp = true positives, fp = false positive, tn = true negative and fn = false negative.

- **Results:**

The experiment was carried out for two possible cases, VSM classifier and VSM Classifier with Sentiment analysis. Also, two weighting schemes were also used for the experiment, first tf.idf and then only idf.

The results are as follows :

Precision	idf	tf.idf
VSM Classifier	32.1 %	38.2 %
VSM Classifier with Sentiment Analysis	49.5 %	38.9 %

Table 1: Prediction Models' Performance

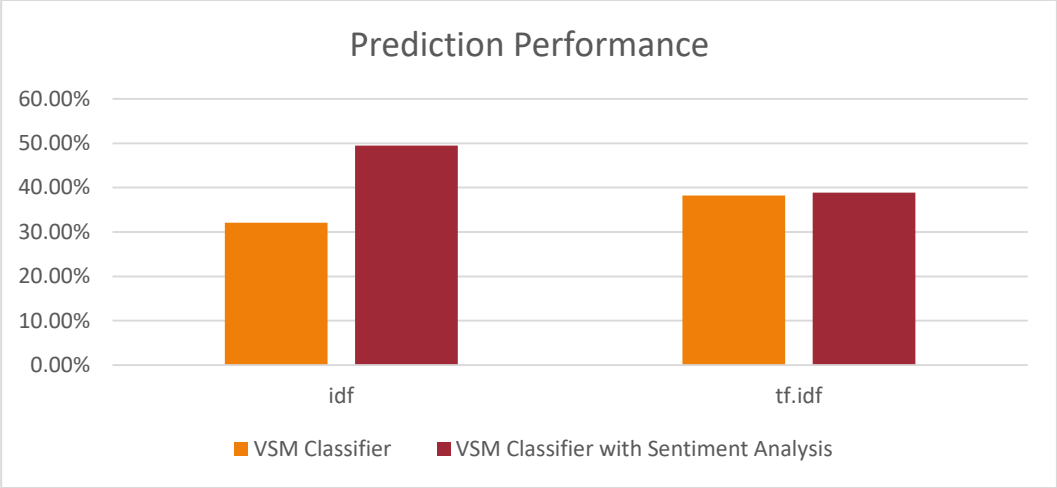

Fig 1: Prediction Models' Performance

On addition of Sentiment Analysis to VSM Classifier, an increase in prediction percentage is observed. If tf.idf weighting scheme is used an increase of only 2% is observed whereas if idf weighting scheme is used, an increase of almost 49.5 % is observed.

- **Limitations and Improvements:**

The prediction percentage as given by the VSM Classifier as well as the VSM Classifier with sentiment analysis model is only around 50 %.

With the help of different models of information retrieval, like

- Probabilistic Model of Information Retrieval
- Statistical Language Model

the prediction performance can be improved significantly.

Further, the sentiment analysis model can be enhanced with cutting edge Natural Language Processing (NLP) procedures.

- **Conclusion:**

This paper tries to identify the problem of predicting ratings based on comments made by users by using a Classifier based on Information Retrieval and the Vector Space Model. The effect of integrating a Sentiment Analysis model is also investigated. The outcome is that VSM based text classifier is improved with the addition of a sentiment analysis model at predicting ratings from the comments.

- **Reference:**

Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis, Eissa M. Alshari, Azreen Azman, Norwati Mustapha, Shyamala A/p C Doraisamy and Mostafa Alksher, IBB University, Yemen, 2016 Third International Conference on Information Retrieval and Knowledge Management