JUNLP at IJCNLP-2017 Task 3: A Rank Prediction Model for Review Opinion Diversification

Monalisa Dey, Anupam Mondal, Dipankar Das
Jadavpur University, Kolkata, India
monalisa.dey.21@gmail.com,
anupam@sentic.net, ddas@cse.jdvu.ac.in

Abstract

IJCNLP-17 Review Opinion Diversification (RevOpiD-2017) task has been designed for ranking the top-k reviews of a product from a set of reviews, which assists in identifying a summarized output to express the opinion of the entire review set. The task is divided into three independent subtasks as subtask-A, subtask-B, and subtask-C. Each of these three subtasks selects the top-k reviews based on helpfulness, representativeness, and exhaustiveness of the opinions expressed in the review set individually. In order to develop the modules and predict the rank of reviews for all three subtasks, we have employed two well-known supervised classifiers namely, Naïve Bayes and Logistic Regression on the top of several extracted features such as the number of nouns, number of verbs, and number of sentiment words etc from the provided datasets. Finally, the organizers have helped to validate the predicted outputs for all three subtasks by using their evaluation metrics. The metrics provide the scores of list size 5 as (0.80 (mth)) for subtask-A, (0.86 (cos), 0.87 (cos_d), 0.71 (cpr), 4.98 (a-dcg), and 556.94 (wt)) for subtask B, and (10.94 (unwt) and 0.67 (recall)) for subtask C individually.

1 Introduction

Review opinion diversification shared task aims to produce top-k reviews for each product from a set of reviews, so that the selected top-k reviews act as a summary of all the opinions expressed in the reviews set. The three independent subtasks incorporate three different ways of selecting the top-k reviews, based on helpfulness, representativeness, and exhaustiveness of the opinions expressed in the review set. The helpfulness refers to the usefulness rating of reviews. Representativeness indicates the popular perspectives expressed in the corpus, whereas exhaustiveness shows the opinion based coverage of reviews of the products (Singh et al., b).

In order to rank and identify the top-k reviews for all the subtasks, we have designed three isolated modules using two well-known machine learning classifiers as Naïve Bayes and Logistic Regression on the top of our extracted features such as number of nouns, verbs, and sentiment etc.

These modules help to resolve the following challenges to identify the top-k reviews of products for the corpus, which is presented as a contribution of the paper.

A. Dataset collection for each subtasks: To the process, the organizers have provided two datasets as training and development [1].

B. Module building for all three subtasks: In order to build the prediction modules, we have extracted various features such as number of nouns, number of verbs, number of negation words, and sentiment etc. from the provided datasets. Thereafter, these features are applied on Naïve Bayes and Logistic Regression classifiers to learn and predict the score of reviews based on helpfulness, representativeness, and exhaustiveness. The predicted scores assist in identifying the top-k reviews of the products from the corpus.

C. Evaluation of the proposed module for all three subtasks: To evaluate, we have processed the test dataset provided by the organizers on the proposed modules and obtained the results for all the three subtasks individually. Thereafter, these results are applied on the evaluation metrics offered

[1] https://sites.google.com/itbhu.ac.in/revopid-2017/data

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 138–142,
Taipei, Taiwan, November 27 – December 1, 2017.

by the organizers to validate all the modules.

The proposed modules help to design various opinion-based diversification applications along with summarization (Krestel and Dokoohaki, 2011; Kacimi and Gamper, 2011; Krestel and Dokoohaki, 2015; Dey,). In the following sections, we have discussed the contribution of the paper as proposed modules and evaluation techniques in details.

2 Proposed Modules

In the present work, we have designed three modules to attempt subtasks A, B, and C according to the properties as usefulness, representativeness, and exhaustiveness. These modules help to identify the top-k reviews against their predicted rank. So, we have used two well-known classifiers namely Naïve Bayes and Logistic Regression in the presence of extracted features from the datasets. The features have been diversified based on the above-mentioned nature and type of the subtasks. The following subsections discuss the data collection, feature extraction, and module building steps in details.

2.1 Data Collection

A well defined dataset is very important to develop any information extraction system. To the process, the organizers provided three datasets namely development, training, and test, which they have collected from Amazon SNAP Review dataset. Thereafter, the organizers have annotated the development and test datasets. The annotated dataset sample of a review contains various features namely, ID of the reviewer, ID of the product, name of the reviewer, helpfulness rating of the review, review text, rating of the product, summary of the review, and time of the review. The development dataset is used for learning the three modules, whereas test dataset is applied for predicting and evaluating the top-k reviews of each product.

2.2 Subtasks Description and Feature Extraction

Subtask-A: Subtask-A aims to produce a ranked list of k reviews based on its predicted usefulness while simultaneously trying to reduce the redundancy among the ranked list.

In the given data, the usefulness rating feature is a user-collected field. We have observed from the development and training datasets that certain linguistic features play a major role in determining the usefulness of a review. Keeping this in mind, we have extracted various features namely, number of words, number of stop words, number of bi-grams, number of trigrams, and tf-idf from the datasets. In order to extract these features, we have written few python (python 2.7) scripts using various packages such as nltk [2].

Subtask-B: The subtask focuses on producing a ranked list of k reviews so as to maximize representativeness of the ranked list. The ranked list of reviews should summarize the opinions expressed in the reviews, both diverse and novel.

To achieve this, we have studied the dataset carefully and observed that the lexical features of the review are presented as an important part in identifying an ideal representation covering popular perspectives. Hence, we have used the features extracted for subtask-A along with three additional features namely, number of verbs, number of nouns, and number of adjectives from the datasets to prepare the final feature set for subtask-B. Number of nouns and verbs help to identify the important linguistic keywords from the reviews, whereas, number of adjectives assist in recognizing useful sentiment keywords.

Subtask-C: Subtask-C emphasizes on producing a ranked list so as to include the majority of opinions regarding the product. The correctness of the list is judged on the basis of how exhaustively the list covers all the opinions. It is to be noted that the ranked list should be the best in expressing all forms of opinions.

In order to achieve this objective, we have chosen to observe the sentiments expressed in each opinion. We have also perceived that both positive and negative opinions have to be included in order to increase the opinion coverage. Besides, linguistic features also take part in determining all forms of viewpoints. Keeping all these observations in mind, we have prepared the feature set by adding few sentiment features namely, number of sentiment words, number of negations, number of positive words, and number of negative words along with the mentioned linguistic features of subtask-B. The sentiment features have been extracted using SentiWordNet [3] and SenticNet [4] re-

[2] www.nltk.org/

[3] http://sentiwordnet.isti.cnr.it/s

[4] http://sentic.net/

sources (Esuli and Sebastiani, 2006; Cambria et al., 2016).

For example, the following review has been labeled with the sentiment features such as number of sentiment words (6), number of negations (2), number of positive words (5), and number of negative words (1).

"It arrived quickly, looks sturdy and doesn't take too much room in the trunk, but I haven't needed it yet, so only 4 stars."

2.3 Modules Building

In order to predict the rank for all the three subtasks, we have applied two conventional supervised machine learning classifiers viz. Naïve Bayes and Logistic Regression. These classifiers have been learned using the extracted features as mentioned in the previous subsections. Thereafter, to predict the final rank for the reviews of the products, we have used the test dataset provided by the organizers. To obtain a single predicted output as rank for each review, we have calculated the average of both of the models predicted scores. The following steps illustrate the overview of the modeling building parts.

Step-1: The development dataset supplied by the organizers has been processed with our written python scripts (python version 2.7) and few sentiment resources to extract various features such as number of nouns, number verbs, number of negation words, and sentiment words etc for all the subtasks.

Step-2: The extracted features are distributed into three segments based on the nature of the subtasks namely helpfulness, representativeness, and exhaustiveness.

Step-3: Thereafter, the segments are processed with the Naïve Bayes and Logistic Regression classifiers to develop all the three modules consequently.

Step-4: The test dataset provided by the organizers is applied on the proposed modules individually to predict the rank of reviews.

Step-5: The predicted ranks help to identify top-k (the value of k decided by the organizers as 5 and 10) reviews of each product from the dataset.

The following section describes the overall evaluation process for all the subtasks.

3 Evaluation

In order to evaluate the output of the proposed modules as top-*k* reviews from the given reviews set for all three subtasks, we have taken help of the organizers provided evaluation metrics (Singh et al., a). The metrics are presented for all the three subtasks. Subtask-A has been evaluated using more than half's (mth), whereas Subtask-B is validated through cosine similarity (cos), discounted cosine similarity (cos_d), cumulative proportionality (cpr), alpha-DCG (a-dcg), and weighted relevance (wt) metrics. On the other hand, unweighted relevance (unwt) and recall metrics are applied to evaluate Subtask-C. The following subsections are discussed about the output of the designed modules for each subtasks in details.

3.1 Validation of Subtask-A

The applied mth metric refers the fraction of reviews included with more than half votes in favour. Hence, they have calculated *Upvotes*, users who found the review helpful and *Downvotes*, users who didn't find the review helpful to find the favour as yes, no, and not counted. The total number of yes favours and combination of yes and no favours help to calculate the mth as shown in Equation 1.

$$mth = \frac{yes}{yes + no}, \tag{1}$$

where yes and no represent the total number of yes and no favours respectively.

The equation assists in measuring the mth score of our proposed module for two different files with list size 5 and list size 10 (Singh et al., b). Table 1 shows a comparative study between our module (JUNLP) and other modules of participants of this shared task.

3.2 Validation of Subtask-B

Subtask-B has been evaluated using five different metrics viz. cosine similarity (cos), discounted cosine similarity (cos_d), cumulative proportionality (cpr), alpha-DCG (a-dcg), and weighted relevance (wt) [5]. Table 2 and Table 3 indicate the mentioned metrics scores for our proposed module

[5]https://sites.google.com/itbhu.ac.in/revopid-201fsu7/evaluation

Participating Groups	**List size 5**	**List size 10**
CYUT1	0.71	0.76
CYUT2	0.84	0.86
CYUT3	0.70	0.75
JUNLP	**0.80**	**0.84**
FAAD1	0.78	0.81
FAAD2	0.78	0.84
FAAD3	0.78	0.83

Table 1: A comparative study between all participants of subtask-A of this shared task for two different files with list size 5 and 10.

and a comparative study between all the submitted module of this subtask.

Metrics	**List size 5**	**List size 10**
cos	0.86	0.90
cos_d	0.87	0.91
cpr	0.71	0.68
a-dcg	4.98	5.71
wt	556.94	1384.6

Table 2: The evaluation output of our proposed module (JUNLP) for subtask-B.

	JUNLP		**BASE_R**	
Metrics	**List size 5**	**List size 10**	**List size 5**	**List size 10**
cos	0.86	0.90	0.84	-
cos_d	0.87	0.91	0.84	-
cpr	0.71	0.68	0.74	-
a-dcg	4.98	5.71	4.53	-
wt	556.94	1384.6	533.41	-

Table 3: A comparative study between all the submitted modules for subtask-B of this shared task.

3.3 Validation of Subtask-C

Another two metrics as unweighted relevance (unwt) and recall are used to validate the output of subtask-C. Unweighted relevance indicates a discounted sum of number of opinions present in the ranked list, whereas recall is the fraction of the relevant opinions that are successfully retrieved by the ranking. The output of the proposed module for subtask-C is presented in Table 4.

Finally, we can conclude that our proposed modules provide noticeable scores for all three subtasks as compared to other participants.

Metrics	**List size 5**	**List size 10**
unwt	10.94	28.93
recall	0.67	0.85

Table 4: Evaluation output of our proposed module (JUNLP) for subtask-C.

4 Conclusion and Future Scopes

This paper presents a rank prediction model for review opinion diversification to IJCNLP-2017 RevOpiD shared task. The task is distributed into three subtasks viz. helpfulness, representativeness, and exhaustiveness based ranking of the product reviews. We have developed three isolated modules for the subtasks individually. Two well-known machine learning classifiers namely Naïve Bayes and Logistic Regression have been applied on the extracted features to design these modules. We are able to obtain noticeable outputs using evaluation metrics provided by the organizers for all the proposed modules. Finally, the paper presents comparative studies between all the submitted systems and our system for all the subtasks. In future, we will attempt to improve the accuracy of our proposed modules by incorporating more fine grained features.

References

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn W Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING*, pages 2666–2677.

Monalisa Dey. Ntcir-12 mobileclick: Sense-based ranking and summarization of english queries.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.

Mouna Kacimi and Johann Gamper. 2011. Diversifying search results of controversial queries. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 93–98. ACM.

Ralf Krestel and Nima Dokoohaki. 2011. Diversifying product review rankings: Getting the full picture. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 138–145. IEEE Computer Society.

Ralf Krestel and Nima Dokoohaki. 2015. Diversifying customer review rankings. *Neural Networks*, 66:36–45.

Anil Kumar Singh, Avijit Thawani, Anubhav Gupta, and Rajesh Kumar Mundotiya. Evaluating opinion summarization in ranking. In *Proceeding of the 13th Asia Information Retrieval Societies Conference (AIRS 2017)*, November.

Anil Kumar Singh, Avijit Thawani, Mayank Panchal, Anubhav Gupta, and Julian McAuley. Overview of the ijcnlp-2017 shared task on review opinion diversification (revopid-2017). In *Proceedings of the IJCNLP-2017 Shared Tasks*, December.