# IJCNLP-2017 Task 3: Review Opinion Diversification (RevOpiD-2017)

**Anil Kumar Singh, Avijit Thawani**
**Mayank Panchal, Anubhav Gupta**
IIT (BHU), Varanasi, India

**Julian McAuley**
University of California, San Diego

## Abstract

Unlike Entity Disambiguation in web search results, Opinion Disambiguation is a relatively unexplored topic. RevOpiD shared task at IJCNLP-2107 aimed to attract attention towards this research problem. In this paper, we summarize the first run of this task and introduce a new dataset that we have annotated for the purpose of evaluating Opinion Mining, Summarization and Disambiguation methods.

## 1 Introduction

In the famous Asch Conformity experiment, individuals were first shown a line segment on a card. Next, they were shown another card with 3 line segments (with a significant difference in length) and were asked to decide which of the 3 matched the length of the previously shown line. The same task was then to be performed in the presence of a group of 8 people (where the remaining 7 were confederates/actors, all of whom were instructed beforehand to give the wrong answer). The error rate soared from a bare 1 percent in the case the subject was alone, to 36.8 percent when the people around expressed the wrong perception (Asch and Guetzkow, 1951). This goes to show how heavily can others' opinions influence our own. With the ever growing influence of sources of opinion today, the need to regulate them is also at an all time high. Documents in the form of social media posts, web blogs, biased or fake news articles, tweets and product reviews can be listed as the primary sources of opinionated information one encounters on a daily basis. Vidulich et. al (Vidulich and Kaiman, 1961) also reported similar results in experiments with the sources of conformity. They found that dogmatists are influenced by the status of the source of information.

The domains of Search Result Ranking and Document Summarization then possess a great potential (and bear a great responsibility) in influencing popular opinion about a target entity. For example, if on searching for 'iPhone reviews', we see results (ranked by, say, PageRank) that coincidentally happen to be against the product, then one might form a perception of the general opinion around the world regarding the smartphone. This perception may or may not be in line with the original composition of the opinion worldwide. What, then, should be the basis of document ranking in Information Retrieval methods?

To delve deeper into addressing this problem, we chose to limit ourselves to a single type of documents: Product Reviews. The reason behind this choice is manifold: Product Reviews are concise, targeted, opinionated (though sometimes descriptive and sometimes objective), diverse (in terms of the category of product), readily available as datasets, and easily comprehended (which makes annotating such data relatively easier). Besides, finding a diverse subset of product review documents (in terms of opinions) provides a good application, which might be of commercial interest to e-commerce websites.

For a product with several reviews, it can get cumbersome for a user to browse through them all. According to an internet source, 90 percent of consumers form an opinion by reading at most 10 reviews, while 68 percent form their opinion after reading just 1-6 reviews.[1] It leads to a natural curiosity into the manner in which reviews are ordered. Order by date (most recent reviews first), order by upvotes (reviews voted 'helpful' the most are ranked first), group by words (show only those reviews which contain specific words, eg. 'battery'), group by stance (segregate reviews

[1] https://www.brightlocal.com/learn/local-consumer-review-survey/



*Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks*, pages 17–25,
Taipei, Taiwan, November 27 – December 1, 2017. 

into positive and negative), group by stars (filter reviews which gave a certain number of stars to the product) are some of the techniques used in sorting and ranking of online customer reviews.

However, only the last two of these take into account the difference of opinions in reviews. And not even these take into account the overall opinion about the product. What we propose is a ranked list that aims to represent a gist of opinions of the whole set of reviews (for any given product). To this end, we will present a novel dataset that can be used as a benchmark for evaluating such a ranked list in Section 3. We will also summarize the details of RevOpiD 2017, the first run of Review Opinion Diversification shared task in Section 4.

## 2 Related Work

Many researchers have undertaken the study of opinion diversity, but most exhibit limited scope owing to the absence of a standard dataset among the community. The Blog Track Opinion Finding Task (TREC 6-8) has a favourable corpus, and was initially meant to judge systems on their Reranking approach on web search results, based on opinion diversity.

The Aspect Based Sentiment Analysis task at SemEval 2014-2016 (Pontiki et al., 2014) was an initiative towards the objective evaluation of sentiment expressed in product reviews. In a wide enough corpora of 39 datasets, ranging across 7 domains and 8 languages, the task was to identify target entity and pick the attribute commented upon (from a list of attributes already provided to annotators).

Our aim differs slightly in that we reward systems which ultimately produce an opinion diversified (and representative) ranking of a subset of the review corpora. The motivation for this statement bases itself on two targeted benefits:

1. Due to absence of an inventory of aspects or opinions for the participants to 'identify', the systems must mine new 'aspects' that vary enormously for different products. Thus, vague aspects in the form of topics modelled will be rewarded equivalently to another approach that, say, manages to match exact lexicons to the subtopics retrieved.

2. We avoid evaluating the opinions on the opinions mined since the number of opinions expressed is a subjective choice made by annotators in the labelling process. For instance, if one annotator suggests having 'affordable' and 'worth the money' as two different opinions whereas a system assumes both to express the same opinion, it may still perform well on diversifying the ranked list. Hence our evaluation on the final ranked list prevents systems from over-fitting on the opinions mined.

Despite the limitations in previous opinion mining evaluations, a recurring and fundamental feature in most of these methodologies is the identification of nuggets (in summarization jargon) or subtopics (in indexing terminology) or attributes (in product reviews); and their subsequent application in having a fine-grained view at the relevance contained in a document. In our pursuit of a tested and suitable data collection, we observed the small-scale attempts at similar data annotation (Marcheggiani et al., 2014) (Täckström and McDonald, 2011) (a few tens of reviews at most, for evaluation of their own opinion mining and discourse analysis systems respectively). The most well known among these is the 'Mining and Summarizing Customer Reviews' paper by Bing et. al. (Hu and Liu, 2004). The experimentation in this publication is based on a compilation of the first 100 reviews of 5 products from Amazon.com and cnet.com. Initially, there were 9 and then 3 more products were added in subsequent years (Ding et al., 2008) (Liu et al., 2015). These reviews were sentence-wise annotated with the following:

1. feature on which opinion is expressed, if any

2. orientation of opinion (+ or -)

3. opinion strength (on a scale of 1 to 3)

An example of annotation by the human taggers (the authors of the paper themselves) for a Digital Camera is: "affordability[+3]while , there are flaws with the machine , the xtra gets five stars because of its affordability ."

## 3 Dataset

The dataset labelled by Bing et. al. is created through a fairly suitable and scalable annotation procedure, despite the inherent flaws associated with subjectivity of human annotation. A

| Sterling Silver Cubic Zirconia Eternity Ring | |
|---|---|
| **Product Reviews** | **Rating** |
| 1. Alex Date : 24/07/2016 | |
| This ring is pretty, it can go good with another ring. It narrow and the stone size is small by it self. It would be a good thumb ring. Again nice ring that does not have alot of bling. | 3.0/5.0 |
| 2. Bran Date : 20/07/2016 | |
| The ring was a gift and my daughter loved it!!! It is very sparkly and fit just right! I would highly recommend this product. | 4.0/5.0 |
| 3. Chau Date : 14/07/2016 | |
| **This ring is amazing for the price**. It doesn't turn my finger white, and the sizing is great. It's just a little bling that isn't too flashy. I wear it as a thumb ring. I think it's really pretty, and very sparkly. | 3.0/5.0 |
| 4. Dany Date : 29/06/2016 | |
| This sterling ring is not too wide, it has a nice touch with the CZ all the way around, making it easier to wear for my wife, because she doesn't worry about it spinning and cutting into the fingers to the side. The CZ stones are recessed a bit making it pretty smooth. | 1.0/5.0 |

Table 1: A Sample Ranking of Product Reviews

few drawbacks are yet to be addressed before we present our Opinion Labelling procedure:

1. Bing et. al. aimed to mine features and opinions from review texts, and hence it is justifiable to practice sentence-wise labelling. On the other hand, for evaluation of opinion diversity in reviews (or any document), labelling of each statement is less of a benefit and very time consuming.

2. The referred dataset contained 96 unique features for a total of 95 reviews (product: Digital Camera 2). Such an exhaustive labelling is again detrimental to the annotation efforts, and is of limited benefits. A reasoning behind this can be observed from the way commercial websites continue to sort their reviews. TripAdvisor, for instance, uses a common set of just 6 attributes: 'Location', 'Service'...

Note that identification of opinions on a per-product basis is a key point of the procedure described in this paper.

### 3.1 Labelling Procedure

Having established our primary objectives behind the need for a opinion-labelled dataset, we now propose our opinion labelling procedure. Labelling process can be broken down into 2 steps. Note that this procedure is to be iterated for each product individually.

1. Make an opinion list, i.e., a set of popular opinions recurrently occurring in the reviews.

2. Make an opinion matrix. The opinion-document matrix (or simply the opinion matrix) is a tabular output of the labelling process, with each row corresponding to a review and each column corresponding to an opinion from the opinion list of the product.

Due to the space constraints, we avoid full textual description and complete specification of guidelines for the dataset. We proceed to show a sample Opinion List (Table 2) and a sample Opinion-Document matrix (Table 3).

### 3.2 Proportion

Our opinion annotated dataset is derived from a subset of Amazon SNAP online reviews dataset (McAuley and Leskovec, 2013). The original SNAP dataset contains more than 34 million reviews spanning over 2 million products. 85 products were chosen from among these, spanning 12 categories, and were labelled with opinions. The number of reviews per product is shown in Table 4 and the number of opinions taken (as considered by annotators) for each product are shown in Table 5. The products in both these tables have been grouped by their category. Eg. Office category has 6 products which are included in our dataset.

| Opinion List |
|---|
| Realistic look |
| Good deal |
| Thumb ring replacement |
| Preference of sizes |
| Good for gifts |
| Matches with jewelry collection |
| Dainty and sparkly ring |
| Comfortable fit |
| Quality product |
| Long lasting and durable |
| Great substitute for wedding ring |
| Cleans easy |
| Not upto the pic |
| Looks expensive |
| Stones are small |
| 'Made in China' on the interior looks bad |
| Stones fallen out |
| Not much sparkly |

Table 2: Opinion List for "Sterling Silver Cubic Zirconia Eternity Ring"

| Opinion Matrix | | | | | |
|---|---|---|---|---|---|
| | Realistic Look | Good deal | Many sizes | Good for gifts | Sparkly ring |
| Review1 | X | X | | | X |
| Review2 | | X | | | |
| Review3 | | | X | X | |
| Review4 | X | | | | X |
| Review5 | X | X | X | | |
| Review6 | X | | | X | |
| Review7 | | | | X | X |
| Review8 | X | | X | | |
| Review9 | | | | | X |
| Review10 | | X | X | X | |
| Overall | 5 | 4 | 5 | 4 | 4 |

Table 3: Opinion Matrix for "Sterling Silver Cubic Zirconia Eternity Ring"

### 3.3 Inter Annotator Confidence

Our proposed evaluation framework relies heavily on the labelling procedure described above, which in turn has the major factor of human subjectivity. What one annotator deems as an opinion (as expressed in a certain number of reviews for a product) might not seem significant enough for another annotator. Thus inter annotator agreement studies are crucial for judging our dataset's reliability. We conducted an experiment asking 5 of our annotators to annotate a single product's review files (only the first 25 reviews). Since opinion lists are not marked 0s or 1s but contain natural language (opinions in the form of text), it is difficult to measure their agreement objectively. Instead, we checked the inter-annotator confidence on whether specific opinions occur in a given review or not. For every pair of annotators A and B, whose inter annotator agreement is to be calculated, we manually select certain opinions from O1 (opinion list of A) which have more or less equivalent opinions in O2 (opinion list of B). Let this set be called O3. Thereafter, presence or absence of opinion $o_i$ in a review $r_i$ in opinion matrix M1 (matrix of A) is compared with that in M2 (matrix of B).

Cohen's Kappa inter-rater agreement (Fleiss and Cohen, 1973) for different pairs of annotators is summarized in Table 6. For example annotators A1 and A2 show a Cohen's Kappa coefficient of 0.77 for the commonly occurring opinion "Realistic look". Some blank cells exist (for example, in A1-A3 and A2-A3 under 'Moderate') since not all opinions occur in the opinion lists of all annotators.

## 4 RevOpiD-2017

RevOpiD-2017 is a part of the 8th International Joint Conference on Natural Language Processing (November 27 to December 1, 2017) at Taipei, Taiwan. The shared task consists of three independent subtasks. Participating systems are required to produce a top-$k$ summarized ranking of reviews (one ranked list for each product for a given subtask) from amongst the given set of reviews. The redundancy of opinions expressed in the review corpus must be minimised, along with maximisation of a certain property. This property can be one of the following (one property corresponds to one subtask):

1. usefulness rating of the review (Subtask A)
2. representativeness of the overall corpus of reviews (Subtask B)
3. exhaustiveness of opinions expressed (Subtask C)

Some Definitions:

1. Review: Review text and any other relevant metadata as may be considered necessary to be used by the participating system, from the given data.

| Reviews per Product | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Baby | 133 | 113 | 111 | 103 | 132 | 124 | 100 | 107 | 125 | 111 |
| Automotive | 149 | 147 | 150 | 110 | 123 | 103 | 101 | 105 | 150 | |
| Health | 112 | 114 | 104 | 101 | 127 | 120 | 103 | 126 | 107 | |
| Grocery | 114 | 118 | 147 | 117 | 122 | 115 | 100 | 146 | | |
| PetSupplies | 105 | 132 | 100 | 128 | 120 | 146 | 105 | | | |
| Beauty | 137 | 143 | 123 | 137 | 109 | 102 | 102 | | | |
| PatioLawn | 143 | 115 | 109 | 104 | 105 | 119 | 150 | | | |
| Office | 135 | 124 | 131 | 119 | 101 | 131 | | | | |
| ToolsHome | 123 | 99 | 146 | 138 | 135 | 131 | | | | |
| DigitalMusic | 130 | 129 | 102 | 125 | 142 | 102 | | | | |
| VideoGames | 117 | 108 | 108 | 101 | 116 | | | | | |
| ToysGames | 114 | 126 | 138 | 149 | 111 | | | | | |

Table 4: Reviews per Product

| Opinions per Product | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Baby | 19 | 27 | 23 | 27 | 17 | 12 | 20 | 20 | 16 | 23 |
| Automotive | 22 | 23 | 23 | 14 | 23 | 13 | 17 | 18 | 23 | |
| Health | 22 | 26 | 23 | 20 | 22 | 39 | 11 | 21 | 18 | |
| Grocery | 21 | 26 | 14 | 21 | 27 | 11 | 17 | 22 | | |
| PetSupplies | 17 | 27 | 12 | 16 | 22 | 15 | 21 | | | |
| Beauty | 22 | 26 | 13 | 16 | 23 | 18 | 20 | | | |
| PatioLawn | 22 | 26 | 17 | 30 | 13 | 19 | 26 | | | |
| Office | 19 | 27 | 18 | 15 | 11 | 18 | | | | |
| ToolsHome | 21 | 18 | 25 | 15 | 29 | 21 | | | | |
| DigitalMusic | 25 | 31 | 18 | 22 | 15 | 18 | | | | |
| VideoGames | 20 | 24 | 11 | 27 | 28 | | | | | |
| ToysGames | 19 | 24 | 32 | 14 | 16 | | | | | |

Table 5: Opinions per Product

2. Corpus: All the reviews for a given product.

3. Feature: A ratable aspect of the product.

4. Opinion: An ordered pair of an aspect and sentiment (for that aspect) in any review.

For the purpose of RevOpiD 2017, our derived dataset was split into three parts:

1. Training Data: was the same as the SNAP dataset, except it being a subset of the latter. Statistics of the training data has been shown in Table 7.

2. Development Data: contained annotated opinion matrices along with the text review files for 30 products. These matrices were used by an evaluation script to measure the performance of participating systems in Subtasks B and C (for representativeness and exhuastiveness).

3. Test Data: The test data contained the review text files alone (also devoid of usefulness scores) of 50 products. The opinion matrices were withheld by us to evaluate final scores based on this test data.

### 4.1 Task Description

#### 4.1.1 Subtask A (Usefulness Ranking)

Usefulness rating is a user-collected field in the provided training dataset. Given a corpus of reviews for a particular product, the goal is to rank the top-$k$ of them, according to predicted usefulness rating, while simultaneously penalizing redundancy among the ranked list of reviews. An essential subsection of this task obviously includes predicting the usefulness rating for a particular review.

| Kappa Inter Rater Agreement (on opinion matrix) scores for 3 of our annotators | | | |
|---|---|---|---|
| | A1-A2 | A2-A3 | A3-A1 |
| Looks Real | 0.77 | 0.77 | 0.61 |
| Perfect Fit | 0.43 | 0.38 | 0.29 |
| Moderate | 1.0 | | |
| Pretty | 0.30 | | |
| Light Weight | 1.0 | | |
| Good Deal | | | 0.34 |
| Different from Image | 0.64 | | |
| Looks Cheap | 0.0 | | |
| Affordable | | 0.66 | |
| Alternate to wedding ring | | 0.78 | |
| Not bright | | | 1.0 |
| No Maintenance | | -0.05 | |
| Looks Expensive | 1.0 | 1.0 | 1.0 |
| Alternate to thumb ring | | | 0.62 |
| Good Gift | | | 1.0 |
| Matches with Jwellery | | | 0.64 |
| Cleans Easy | | | 0.0 |
| Overall | 0.59 | 0.59 | 0.61 |

Table 6: Inter Rater Agreement (on opinion matrix) Kappa scores for 3 of our annotators. Product category: Automotive. Number of reviews: 25

.

| Data Statistics | | | |
|---|---|---|---|
| | Products | Reviews | Avg Reviews per Products |
| Automotive | 569 | 172106 | 302 |
| Baby | 1000 | 352231 | 352 |
| Beauty | 1000 | 316536 | 316 |
| Digital music | 468 | 145075 | 309 |
| Grocery | 800 | 293629 | 367 |
| Health | 1000 | 357669 | 357 |
| Office | 1000 | 327556 | 327 |
| Patio lawn | 859 | 263489 | 306 |
| Pet supplies | 1000 | 398658 | 398 |
| Tools home | 1000 | 320162 | 320 |
| Toys games | 1000 | 314634 | 314 |
| Video games | 1000 | 358235 | 358 |

Table 7: Data Statistics

#### 4.1.2 Subtask B (Representativeness Ranking)

Given a corpus of reviews for a particular product, the goal is to rank the top-$k$ of them, so as to maximize representativeness of the ranked list. The ranking should summarize the perspectives expressed in the reviews given as input, incorporating a trade-off between diversity and novelty.

An ideal representation would be one that covers the popular perspectives expressed in the corpus, in proportion to their expression in the corpus (for that product), e.g. if 90 reviews claim that the iPhone cost is low, and 10 reviews claim that it is high, the former perspective should have 90 percent visibility in the final ranking and the latter should have 10 percent (or may even be ignored owing to low popularity) in the final ranking. The ranking should be such that for every i in $1 <= i <= k$, the top i reviews best represent the overall set of reviews for the product. That is,

the #1 review should be the best single review to represent the overall opinion in the corpus; The combination of #1 and #2 reviews should be the best pair of reviews to represent the corpus, and so on.

#### 4.1.3 Subtask C (Exhaustive Coverage Ranking)

Given a corpus of reviews for a particular product, the goal is to rank the top-$k$ of them, so as to include the majority of popular perspectives in the corpus regarding the product, while simultaneously penalizing redundancy among the ranked list of reviews. This is similar to Subtask B, except that:

In Subtask B, the final ranking is judged on the basis of how well the ranked list represents the most popular opinions in the review corpus, in proportion. In Subtask C, the final ranking is judged on the basis of the exhaustive coverage of the opinions in the final ranking. That means, most of the significant (not necessarily all very popular) perspectives should be covered regardless of their proportions of popularity in the review corpus, e.g. if 90 reviews claim that the iPhone cost is low, and 10 reviews claim that it is high, both perspectives should be more or less equally reflected in the final ranked list.

### 4.2 Evaluation

This being the first run of RevOpiD, we experimented with several measures of evaluation (Singh et al.) and 8 of them were shortlisted to study their variations with the system submissions:

1. mth (More than half's): The fraction of reviews included (in submitted ranked list) with more than half votes in favour. In other words, if upvotes on a review be counted as the number of users who found it helpful, and downvotes be counted as the number of users who didn't find it helpful; then the $mth$ count will be incremented by one if upvotes $>$ downvotes.

2. Cosine similarity: Cosine similarity between Overall Vector and Opinion Vector

3. Discounted Cosine similarity: Cosine similarity between Overall Vector and Discounted Opinion Vector

4. Cumulative Proportionality: Based on Saint Lague method, used in Electoral seat allocation. (Dang and Croft, 2012). A ranking S is said to be proportional to the corpus D, or a proportional representation of D, with respect to opinions/aspects T, if and only if the number of documents in S that is relevant to each of the aspects $t_i \in T$ is proportional to its overall popularity $p_i \in D$.

5. $\alpha$-DCG: A measure that rewards novel information (to be covered incrementally in each review) (Clarke et al., 2008).

6. Weighted Relevance: Discounted Cumulative Gain with the relevance of a review given by sum of weights of the opinions covered in the review (weight of an opinion = Number of reviews in which it appears in the whole opinion matrix / corpus size).

7. UnWeighted Relevance: A discounted sum of number of opinions present in the ranked list.

8. Recall: The fraction of opinions/columns covered by the ranking. An opinion is said to be covered if atleast a single 1 appears in that column in the ranked list submission.

### 4.3 Systems

There were 3 participating systems at RevOpiD-2017, namely JUNLP, CYUT and FAAD. Also included in our analysis is the official baseline (Subtasks B and C) [2]. The last row shows the scores obtained for a random submission script averaged over 5 runs.

While CYUT and FAAD have attempted Subtask A alone, JUNLP has submitted runs for each of Subtasks A, B and C.

1. JUNLP (Dey et al.): Instead of posing this as a regression problem, they have modeled it as a classification task where the aim is to identify whether a review is useful or not. They've employed a bi-directional LSTM to represent each review which is used with a softmax layer to predict the usefulness score. First they choose the review with highest usefulness score, then they find its cosine similarity score with rest of the reviews. This is done in order to ensure diversity in the selection of top-$k$ reviews.

[2] https://github.com/shreyansh26/RevOpiD/tree/master

| RevOpiD final scores | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | mth | cos_d | cos | cpr | a-dcg | wt | unwt | recall |
| CYUT 1 | 0.71 | 0.83 | 0.84 | 0.7 | 4.28 | 504.18 | 14.31 | 0.71 |
| CYUT 2 | 0.84 | 0.87 | 0.88 | 0.7 | 5.22 | 575.58 | 17.67 | 0.83 |
| FAAD 1 | 0.78 | 0.86 | 0.87 | 0.49 | 4.27 | 494.03 | 14.04 | 0.76 |
| FAAD 2 | 0.78 | 0.85 | 0.86 | 0.52 | 4.34 | 495.35 | 14.34 | 0.75 |
| FAAD 3 | 0.78 | 0.84 | 0.85 | 0.51 | 4.11 | 486.51 | 13.35 | 0.72 |
| JUNLP A | 0.8 | 0.83 | 0.85 | 0.46 | 4.05 | 475.54 | 13.12 | 0.74 |
| JUNLP B | 0.7 | 0.86 | 0.87 | 0.71 | 4.98 | 556.94 | 16.9 | 0.81 |
| JUNLP C | 0.53 | 0.8 | 0.81 | 0.3 | 3.58 | 390.44 | 10.94 | 0.67 |
| Baseline 0 | 0.64 | 0.84 | 0.84 | 0.74 | 4.53 | 533.41 | 15.33 | 0.73 |
| Baseline 1 | 0.64 | 0.87 | 0.87 | 0.56 | 4.61 | 564.02 | 15.81 | 0.76 |
| Baseline 2 | 0.65 | 0.86 | 0.86 | 0.54 | 4.6 | 566.68 | 15.85 | 0.75 |
| Baseline 3 | 0.63 | 0.86 | 0.87 | 0.56 | 4.6 | 572.27 | 15.97 | 0.75 |
| Expected | 0.61 | 0.79 | 0.81 | 0.11 | 3.4 | 393.07 | 10.45 | 0.64 |

Table 8: System and Baseline Scores

2. CYUT (Wu et al.): This team (with prior work in helpfulness rating prediction of Chinese online reviews) implemented two models using linear regression with two different loss functions: least squares (CYUT 1) and cross entropy (CYUT 2).

3. FAAD (Mishra et al.): Two supervised classifiers (Naıve Bayes and Logistic Regression) are fitted on top of several extracted features such as the number of nouns, number of verbs, and the number of sentiment words etc. from the provided development and training datasets. Three runs (FAAD 1,2,3) vary only in the weightage given to the two classifiers.

4. Baseline: A feature based opinion extraction based on the work of Bing et al. This task is done in three steps:

   (i) Identify the features of the product that customers have expressed opinions on (called opinion features) and rank the features according to their frequencies that they appear in the reviews.

   (ii) For each feature, identify how many customer reviews have positive, negative or neutral opinions. The specific reviews that express these opinions are attached to the feature.

   (iii) Generate an opinion matrix based on these predicted occurrences of opinions and greedily select the best representative and exhaustive rankings.

### 4.4 Results

The results of RevOpiD-2017 have been summarized in Table 8 for the chosen metrics already described above.

Based on the system performances, the feature selection mechanism in CYUT's submission using Cross Entropy loss function proves the leader in Subtask A. JUNLP's submission (representative ranking) outperforms others marginally in Subtask B and Subtask C. Not a lot of improvement is reported over the baseline, therefore there exists a lot of scope for improvement in Subtasks B and C.

## Acknowledgments

This project was assisted upon by several enthusiastic students as well as lab annotators: Shreyansh Singh (for developing the baseline), Shashwat Trivedi, Shivam Arora, Avijeet Diwaker, Divyanshu Gupta, Avi Chawla, Ayush Sharma, Tara Hemaliya, Vandana Singh, Neelam and Rajesh Kumar Mundotiya.

## References

Solomon E Asch and H Guetzkow. 1951. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, pages 222–236.

Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM.

Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 65–74. ACM.

Monalisa Dey, Anupam Mondal, and Dipankar Das. Junlp: Ijcnlp-2017 revopid- a rank prediction model for review opinion diversification. In *Proceedings of the IJCNLP-2017 Shared Tasks*.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Proceeding of the IJCAI*, pages 1291–1297.

Diego Marcheggiani, Oscar Täckström, Andrea Esuli, and Fabrizio Sebastiani. 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *Proceeding of the European Conference on Information Retrieval*, pages 273–285. Springer.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM.

Pruthwik Mishra, Prathyusha Danda, Silpa Kanneganti, and Soujanya Lanka. Ijcnlp-2017 revopid shared task: A bidirectional-lstm approach for review opinion diversification. In *Proceedings of the IJCNLP-2017 Shared Tasks*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of SemEval*, pages 27–35.

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)*.

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceeding of the European Conference on Information Retrieval*, pages 368–374. Springer.

Robert N Vidulich and Ivan P Kaiman. 1961. The effects of information source status and dogmatism upon conformity behavior. *The Journal of Abnormal and Social Psychology*, 63(3):639.

Shih-Hung Wu, Su-Yu Chang, and Liang-Pu Chen. System report of cyut team at revopid-2017 shared task in ijcnlp-2017. In *Proceedings of the IJCNLP-2017 Shared Tasks*.