Polarity Reversal of Online Reviews of Kitchen and Electronic Products

Marvin Kaster **Yogesh Shrestha** **Jan Buchmann**

1 Introduction

The ever-growing use and availability of the world wide web and social media allow more and more people to express their *opinions* towards certain *objects* (e.g. products, companies, people, ideas or political parties) to a (theoretically) earth-spanning audience. Knowing the public opinion towards such an object can be of great use. One of many examples for this is a company which would like to improve its products by analyzing customer reviews about this product.

The sheer number of uttered opinions creates the need for computational analysis methods. *Sentiment analysis* (also known as *opinion mining*) describes a collection of methods which classify texts into a category of (usually positive or negative) sentiment [1]. In the simplest case, the categories are only *positive* and *negative*. These can be extended to express gradual differences (e.g. *slightly* positive or *very* positive), and a neutral category can be added.

Another task in text classification is known as *domain* classification. Here, the goal is to extract or characterize the object that is dealt with in a text (e.g. whether a product review is about a product from the kitchen domain or the electronics domain).

For the present work, several techniques to reverse the sentiment of a product review were implemented and tested. A dataset of 400 reviews labeled with domain information (electronics or kitchen) and sentiment information (positive or negative) and two trained classifiers for sentiment and domain were provided. The classifiers were presented as "black boxes" without information on their inner workings. The aim was to reduce the sentiment classifier accuracy as much as possible while retaining the domain classifier accuracy.

2 Methods

The main idea in all approaches tested was to exchange specific words in a review text to other words of reversed sentiment in order to revert the classifier's prediction. From visual inspection of the texts we identified adjectives and adverbs (and sometimes verbs) as the main carriers of sentiment information, while the nouns carry more domain-related information.

Additionally, "not" was inserted before every verb in all approaches. In the cases where this resulted in two successive "not"-tokens, both were removed.

The first approach used a score generated for every adjective and adverb to decide about their exchange. Information on the syntactic role of a word was taken from the point-of-speech (POS) tag generated with spacy and the 'en_core_web_md' model. The score was generated from the counts of a word in the provided positive and negative review texts, respectively. The scores were calculated specific to the syntactic role of a word (a word could have one score for its use as an adjective and one score for its use as an adverb):

$$s = \log \frac{\frac{c_p+1}{N_p}}{\frac{c_n+1}{N_n}} \tag{1}$$

where s is the score of one particular word, c_p and c_n are the counts of the word in the positive and negative reviews, respectively, and N_p and N_n are the total numbers of words in the positive and negative texts, respectively. A pseudocount of 1 is added to the counts to circumvent problems with counts of 0.

The score was used to exchange adjectives and adverbs in three ways. Approach 1.1: For a given word, the score was reversed (negated). The word with the closest score to the reversed score

and the same syntactic role was found and used to replace the given word. Approach 1.2: The 'polarity' of a given word was determined by checking whether its score is positive or negative. It was replaced by 'good'/'bad' or 'well'/'badly' depending on score and syntactic role. (e.g. an adjective with score 1 was replaced by 'bad') Approach 1.3: Similar to 2, but the words with the highest / lowest scores in the corpus were used for replacement ('amazing'/'disappointed', 'beautifully'/ 'completely').

The second approach uses the spacy wordnet to find the antonyms of adverbs and adjectives which are very decisive to determine to polarity of a review. First, it tries to find the antonym of adverbs and adjectives if there exists any. If the antonym of adjectives and adverbs does not exist in wordnet, it calculates the similarity score of these adjectives and adverbs with respect to two words 'good' and 'bad'. If the similarity score with respect to 'good' is higher than that of 'bad', it replaces the corresponding adjectives and adverbs with the word 'good' and vice versa. If part-of-speeches(POS) 'AUX' and 'VERB' are adjacent to each other, it inserts 'not' between the corresponding tokens. If 'not' already exists between the words which represents POS 'AUX' and 'VERB', it removes 'not'. Removing/Adding 'not' between part-of-speeches has a significant influence on the score of polarity.

3 Discussion

Polarity of an opinion normally falls into three categories: positive, negative, and neutral. But only two categories positive and negative polarity are considered here. There are positive and negative words which are found in opinion texts and play a vital role in distinguishing the polarity (e.g positive words such as good, impressive, beautiful, etc and negative words such as bad, ugly, worst, etc). These words belong mostly either to Part-of-Speech 'ADJ' or 'ADV'. Replacing these words with their antonyms leads to reversing the polarity of the opinion. Sometimes, opinions use negative words which have positive polarity and opinions which uses positive words mostly but has negative polarity. A human being can differentiate it but training machine to differentiate it is a very challenging task. Some verbs have also significant impact on polarity (e.g. hate, love, like, dislike, etc.). Negation of a verb and replacing the positive and negative words at the same time does not change the polarity of the opinion. So, the second approach does not change the polarity of some reviews. Because of the lack of knowledge of how the classifier has been trained ("black box"), it is quite difficult to reverse the polarity of the reviews.

Table 1: Accuracy of classification of the review texts modified with different methods.

Approach	Accuracy (%) (Sentiment / Domain)
1.1	52.00 / 92.50
1.2	50.25 / 90.75
1.3	50.50 / 75.75
2	50.50 / 91.50

We modified the provided review texts using all methods explained in the Methods Section. The modified texts were then classified with the provided classifiers. Classification accuracy values for sentiment and domain are shown in Table 1.

The best-performing variation of the score-based approach is approach 1.2. This might seem as a surprise, because it makes the least use of the given data (the exchange words "good"/"bad" and "well"/"badly" were determined heuristically). However, the used data is the test data. This means that it was not used to train the classifier, and its use can introduce a certain bias, especially due to the small size of the test set. One can assume that the distribution of words is different in the training dataset and the test dataset. One can also assume that the classifier works by 'learning' the correlation between a word and the classification of a text containing the word. The calculation of a score for every word tries to capture this correlation. The correlation between certain words and the classification might be different in the two datasets. The scores calculated from the test data might not always reflect the correlation learned by the classifier. This can be seen in the adverb with the most negative score ("completely"), which is not connotated with any sentiment. Therefore, word exchange based on these scores might not always be optimal to revert the sentiment of a given text. Using heuristically chosen words with commonly known sentiment connotation introduces less bias into the exchange.

There is a notable reduction in the domain classification score in approach 1.3. One explanation for this is that the words used for exchange do not only differ in distribution between the texts with positive and negative sentiment, but also between the texts from the different domains. Assuming the classifiers learn this difference in distribution (the correlation between a word and its classification), the introduction of these words into a text changes the classification results for both classifiers.
In contrast to approaches 1.1 to 1.3, approach 2 makes use of additional data in the form of WordNet antonyms and spacy word vectors to decide on words for exchange. Thereby, this approach is not susceptible to bias in the test data. Nevertheless, the classification accuracy obtained on the texts modified using approach 2 is very similar to the classification accuracies obtained on the texts modified with approaches 1.1 and 1.2.
It would have been interesting to know whether the classifiers consider word order. If it does, the results might be improved by more sophisticated text modification methods. If it does not, the best results are probably obtained by finding the words which have the strongest correlation with the sentiment classification and the weakest correlation with the domain classification in the training data. These words can then be used for replacement. Similarly, words which give "correct" signals for classification can be removed to reduce classification accuracy.

4 Conclusion

All approaches tested work by replacing adjectives and adverbs with words of opposite sentiment and inserting or removing "not" before verbs. Although the exact mechanisms for replacement are different, similar values for classification accuracy are obtained and approach 50%.

References

[1] Bing Liu. Sentiment Analysis and Opinion Mining. In *Synthesis Lectures on Human Language Technologies*. Morgan-Claypool Publishing, 2012.