Hybrid RNN at SemEval-2019 Task 9: Blending Information Sources for Domain-Independent Suggestion Mining

Aysu Ezen-Can
SAS Inst.
aysu.e.can@gmail.com

Ethem F. Can
SAS Inst.
ethfcan@gmail.com

Abstract

Social media has an increasing amount of information that both customers and companies can benefit from. These social media posts can include Tweets or be in the form of vocalization of complements and complaints (e.g., reviews) of a product or service. Researchers have been actively mining this invaluable information source to automatically generate insights. Mining sentiments of customer reviews is an example that has gained momentum due to its potential to gather information that customers are not happy about. Instead of reading millions of reviews, companies prefer sentiment analysis to obtain feedback and to improve their products or services.

In this work, we aim to identify information that companies can act on, or other customers can utilize for making their own experience better. This is different from identifying if reviews of a product or service is negative, positive, or neutral. To that end, we classify sentences of a given review as **suggestion** or **not suggestion** so that readers of the reviews do not have to go through thousands of reviews but instead can focus on actionable items and applicable suggestions. To identify suggestions within reviews, we employ a hybrid approach that utilizes a recurrent neural network (RNN) along with rule-based features to build a domain-independent suggestion mining model. In this way, a model trained on electronics reviews is used to extract suggestions from hotel reviews.

1 Introduction

With the growth of social media usage, the interest in text mining approaches has increased. One task that has gained momentum recently is sentiment analysis where the goal is to determine opinions/emotions from a text input, generally a product or service review. Different approaches have been proposed for sentiment analysis task such as multilingual models to be used with limited data (Can et al., 2018) and sentiment lexicons (Banea et al., 2008). Twitter posts also has been one source of reviews to be mined in terms of sentiment (Pak and Paroubek, 2010; Ezen-Can and Can, 2018; Tellez et al., 2017).

The task of suggestion mining is similar to sentiment analysis in that the input is the same (e.g., customer reviews). However, the output of a sentiment analysis model and a suggestion mining model is different. While sentiment classifiers focus on grouping reviews as positive or negative, suggestion mining models identify the reviews that have suggestions/actionable items/advice to other people/service providers.

In this paper, we present a suggestion mining model that takes product/service reviews and classifies each sentence in a given review as suggestion or not suggestion. To that end, we employ a hybrid LSTM model that utilizes both the textual reviews and features extracted from a rule-based approach.

2 Related Work

For the suggestion mining task, there is not a large body of work in the NLP community. (Brun and Hagege, 2013) use a corpus of reviews of printers made by different manufacturers. Their approach relies on linguistic information such as thesaurus, parser and patterns. (Goldberg et al., 2009) address the task of suggestion mining as a 'wish detection' task and use templates to detect wishes on product reviews and political discussion posts. (Dong et al., 2013) focus on Tweets and classify them as containing suggestion or not by using factorization machines. (Wicaksono and Myaeng, 2013) employed Hidden Markov models with three different sets of features: syntactic, contextual, and sentence informativeness features.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1199–1203
Minneapolis, Minnesota, USA, June 6–7, 2019.

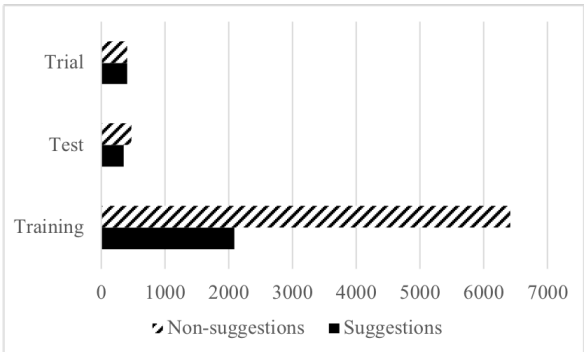

Figure 1: Distribution of classes in the training, trial and test sets.

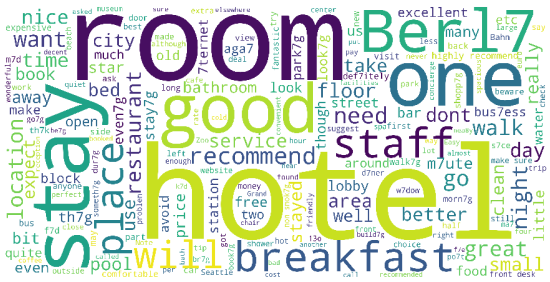

Figure 2: Word cloud of the reviews in the test set.

Recently, (Negi and Buitelaar, 2017) collected a new corpus for suggestion mining (not available at the time of this writing).

Our approach is different from the existing prior work in that we use a hybrid approach where a deep learning model is used in addition to a rule-based technique. The features extracted by rule-based approach are utilized as information sources to an LSTM network where the customer reviews are also fed into as textual input. In this way, we intend to use as many information sources as possible to improve results of a suggestion mining classifier.

3 Corpus

The corpus provided by the Semeval 2019 Suggestion Mining Challenge (Negi et al., 2019) was highly imbalanced as can be seen in Figure 1. There was a total of 8500 reviews, only 2085 of which were suggestions. The test set consisted of 824 observations.

Due to the nature of the challenge, the training set and the test set were from different domains. While the training set contained software/application reviews, test set was collected from hotel reviews. An excerpt from the training set can be seen in Table 1. The word clouds for

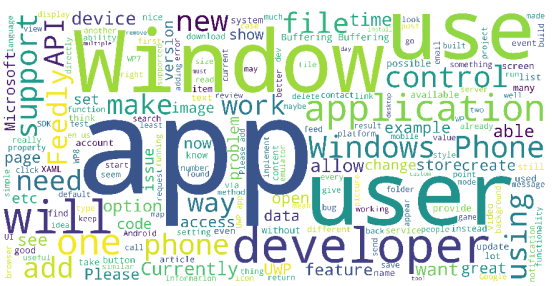

Figure 3: Word cloud of the reviews in the training set.

Review	Class
"I would like to be able to enable WP alerts be forwarded to XBox One when I am near it or manually configured for it."	Suggestion
"When you apply new policies on already existing, especially if it is related to name, all the existing credibility and market is lost."	Non-suggestion
"I find myself having to manually tab out get figures, enter them in."	Non-suggestion
"Possible solution: Route class implements IRoutePath."	Non-suggestion
"Street names color stays black and not being centered."	Non-suggestion

Table 1: Excerpt from the training set.

these two datasets (Figures 2 and 3) show the difference in the most frequently used words.

4 Methodology

In this section, we explain the model architecture used for the task of suggestion mining and the features utilized by the model.

4.1 Features

For suggestion mining, we used two sets of features: rule-based and model-generated from word embeddings. In this section, we describe each of these features.

4.1.1 Rule-Based Features

The rule-based features are extracted from the heuristics used in the baseline system for this challenge. Below are explanations of each of these rule-based feature.

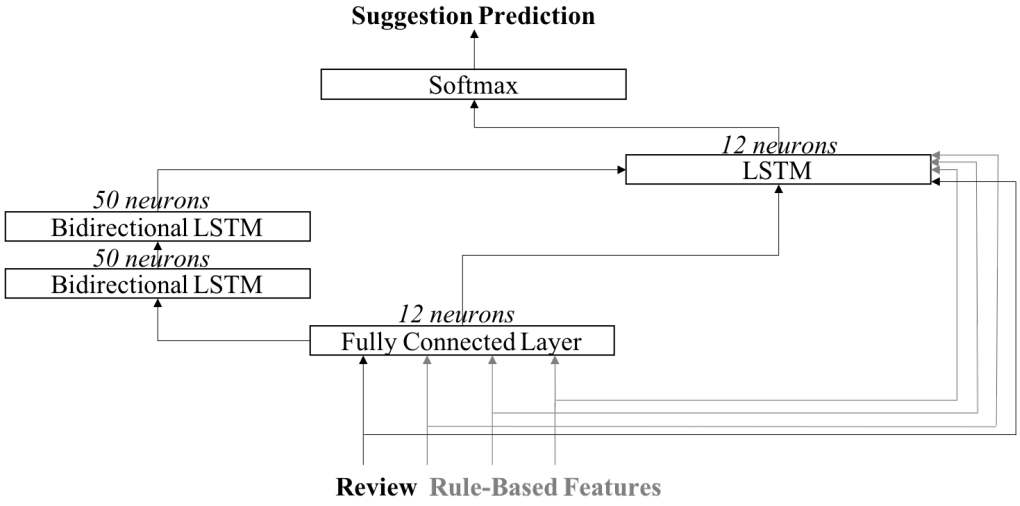

Figure 4: RNN architecture incorporating two different sources of information.

- *Rule-Based Feature 1:* the first rule-based feature is using a pattern matching algorithm based on regular expressions. This heuristic focuses on finding keywords and patterns within the input text such as '.*would \s like.*if.*' and '.*i \s wish.*'. Existence of at least one of these patterns in the review triggers a value of 1 for this feature, 0 otherwise. There are 13 patterns for this heuristic.

- *Rule-Based Feature 2:* the second rule-based feature utilizes keywords without any patterns such as 'suggest', and 'recommend'. Once one of the keywords in the list is present in the given review, the rule flags a 1 value indicating that the review contains a suggestion. There are 17 keywords in total.

- *Rule-Based Feature 3:* the third rule-based feature relies on part of speech tags. There are two part of speech tags that this heuristic is looking for (i.e., MD and VB) to be present in the tagged review to come to the conclusion that the given review is a suggestion.

4.1.2 Word Embeddings

Recurrent neural networks requires a mechanism to convert textual input to numerical vectors to be able to perform computations. To this end, we used pre-trained word embeddings where each word in the embedding table has a vector of size 100. In this study Glove embeddings is used which was trained on Wikipedia 2014 and Gigaword 5 corpora (Pennington et al., 2014).

4.2 RNN Architecture

As part of the RNN architecture, we used a fully-connected layer that takes the rule-based features and the review as the inputs. Then two bidirectional LSTM layers follow for modeling the textual input. Before the softmax layer, an LSTM layer takes the advantage of both learned representations form bidirectional layers and the rule-based features. Figure 4 depicts the architecture of the RNN model.

In the bidirectional layers, we used a dropout of 0.2 and MSRA initialization (He et al., 2015) in all layers. The training set is shuffled randomly before the first epoch. During training, ADAM optimizer (Kingma and Ba, 2014) with gradient clipping is used.

4.3 Ensemble

To fuse the different approaches utilized during the modeling phase, we used an ensemble technique. This technique take the outputs of both the rule-based features and the RNN model. If one of the rule-based features classify the review as a suggestion, the ensemble concludes that the review is a suggestion. If rule-based features classify the review as a non-suggestion and RNN classifies as a suggestion, the overall ensemble labels the observation as a suggestion. Otherwise, a non-suggestion tag is used. It is important to note that, RNN is also incorporating the rule-based features in the model. As can be seen in Figure 4, two main sources of information are fed into the model.

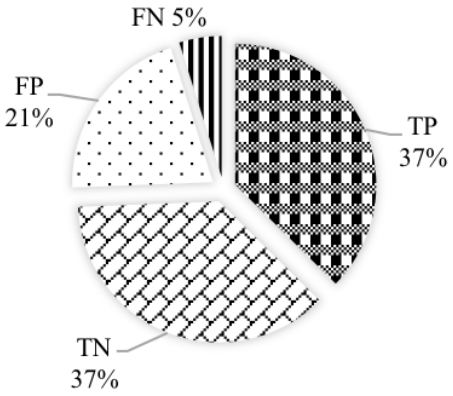

Figure 5: Pie chart showing false / true positives / negatives in the final predictions on the test set.

"and I was woken by the early morning firing up of the local bus service (a courtyard-facing room is essential unless you have industrial strength earplugs.)..."
"Don't eat breakfast in the restaurant, too costly."
"Look around the same area for another hotel."
"Avoid these rooms - it is very clear why they do not have a photograph of them on their web site."

Table 2: Samples from true positives.

5 Results

In this section, we report the results for the test set as well as discussion on the results.

5.1 Experimental Results

For the trial dataset, where the domain was hotel reviews and the majority baseline was 50%, the hybrid approach achieved F1 measure of 77.70%. It is important to note that, trial dataset has not been used to tune or validate the model. With the test dataset, the model obtained 74.49% F1 score where the majority baseline was 57.77%.

5.2 Discussion

Figure 5 shows the ratios of true/false positives and true/false negatives. From our investigation, we found out that the hybrid approach was useful in generalization of the model where the reviews did not have any keywords or patterns defined in the rules. Since RNN used generic pre-trained word embeddings (not specifically trained on either of the domains of the training set or the test set), generalization is expected for RNN. Some examples of such test observations can be seen in Table 2.

"Only one almost useless pillow per person though (think no thicker than a cracker) and no availability of additional bed linen as most other hotels would normally provide."
"Leaving your bedroom window open is not an option as my heavily bitten body will testify!"
"No shampoo provided in the room, Shower Gel dispensers don't work well."
"Put your towel on the wet floor or you will definitely slip."

Table 3: Samples from false positives.

An interesting finding in the results is about false positives. The trend observed in the false positives is that, the reviews that were helping other customers and giving hints *to the customers* rather than *to the service providers* were considered as suggestions by the model. Table 3 shows examples of those reviews where the ground truth considered these reviews as non-suggestions. However, they are suggestions to the receiving end of the service. This finding shows the difficulty of classifying this dataset because suggestions to customers and service providers can both be considered as suggestions (although not labeled as such in the ground truth).

6 Conclusion

Suggestion mining is a crucial task for mining social media data so that companies can focus on services that need improvement. Most of the times, obtaining labeled data in several different domains is not easy. Therefore, in this paper, we focused on domain-independent suggestion mining models where the training set and test set have reviews for different domains. To make our model robust, we utilized a hybrid approach that incorporates both rule-based features and relationships extracted by LSTM from raw text input. Instead of having to decide between rule-based approaches and deep learning, we fused the information sources in two ways. First by using external features in RNN and second by ensembling the result of RNN with rule-based features. By incorporating multiple information sources, we showed that the suggestion mining accuracies outperformed the baseline.

References

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC*, volume 8, pages 2–764.

Caroline Brun and Caroline Hagege. 2013. Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science*, 70(79.7179):5379–62.

Ethem F Can, Aysu Ezen-Can, and Fazli Can. 2018. Multilingual sentiment analysis: An rnn-based framework for limited data. *arXiv preprint arXiv:1806.04511*.

Li Dong, Furu Wei, Yajuan Duan, Xiaohua Liu, Ming Zhou, and Ke Xu. 2013. The automated acquisition of suggestions from tweets. In *AAAI*.

Aysu Ezen-Can and Ethem F Can. 2018. Rnn for affects at semeval-2018 task 1: Formulating affect identification as a binary classification problem. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 162–166.

Andrew B Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sapna Negi and P Buitelaar. 2017. Suggestion mining from opinionated text. *Sentiment Analysis in Social Networks*, pages 129–139.

Sapna Negi, Tobias Daudert, and Paul Buitelaar. 2019. Semeval-2019 task 9: Suggestion mining from online reviews and forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Eric S Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Ranyart R Suárez, and Oscar S Siordia. 2017. A simple approach to multilingual polarity classification in twitter. *Pattern Recognition Letters*, 94:68–74.

Alfan Farizki Wicaksono and Sung-Hyon Myaeng. 2013. Automatic extraction of advice-revealing sentences foradvice mining from online forums. In *Proceedings of the seventh international conference on Knowledge capture*, pages 97–104. ACM.