SAS 4180-2020

Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data

Ethem Can and Aysu Ezen-Can, SAS Institute Inc.

ABSTRACT

Sentiment analysis is a widely studied natural language processing task, whose goal is to determine users' opinions, emotions, and evaluations of a product, entity, or service that they review. One of the biggest challenges for sentiment analysis is that it is highly language-dependent. Word embeddings, sentiment lexicons, and even annotated data are language-specific. Furthermore, optimizing models for each language is very time-consuming and labor-intensive, especially for recurrent neural network (RNN) models. From a resource perspective, it is very challenging to collect data for different languages.

In this paper, we look for an answer to the following research question: Can a sentiment analysis model that is trained on one language be reused for sentiment analysis in other languages where the data are more limited? Our goal is to build a single model in the language that has the largest data set available for the task and reuse that model for languages that have limited resources.

For this purpose, we use reviews in English to train a sentiment analysis model by using recurrent neural networks. We then translate those reviews into other languages and reuse the model to evaluate the sentiments. Experimental results show that our robust approach of training a single model on English-language reviews outperforms the baseline in several different languages.

INTRODUCTION

Steady growth in commercial websites and social media venues has led to easier access to users' reviews. As the amount of data that can be mined for opinion has increased, commercial companies' interests in sentiment analysis has also increased. Sentiment analysis is an important part of understanding user behavior and opinions about products, places, or services.

Sentiment analysis has long been studied by the research community, leading to several sentiment-related resources such as sentiment dictionaries that can be used as features for machine learning models (Banea, Mihalcea, and Wiebe 2008; Inui and Yamamoto 2011; Steinberger et al. 2012; Taboada et al. 2011). These resources help increase the accuracy of sentiment analysis, but they are highly dependent on language and they require researchers to build such resources for every language to process.

Feature engineering constitutes a large part of the model-building phase for most sentiment analysis and emotion detection models (Ortigosa, Martin, and Carro 2014). Determining the correct set of features is a task that requires thorough investigation. Furthermore, these features are highly dependent on language and the particular data set, making it even more challenging to build models for different languages. For example, sentiment and emotion lexicons, as well as pretrained word embeddings, are not completely transferable to other languages. Therefore, modeling efforts must be replicated for every language on which you want to build sentiment classification models. For languages and tasks where the data are limited, extracting these features, building language models, training word embeddings, and creating lexicons are big challenges. In addition to the feature engineering effort, the machine

learning model's parameters also need to be tuned separately for each language in order to obtain optimal results.

In this paper, we take a different approach. We build a reusable sentiment analysis model that does not use any lexicons. Our goal is to evaluate how well a generic model can be used to mine opinions in other languages where data are more limited than the language on which the generic model is trained. To that end, we train a recurrent neural network (RNN) model to predict polarity of reviews. To evaluate the reusability of the sentiment analysis model, we test with non-English data sets. We first translate the test set to English and use the pretrained model to score polarity in the translated text. In this way, our proposed approach eliminates the need to train language-dependent models for their use of sentiment lexicons and word embeddings. Our experiments show that a generalizable sentiment analysis model can be used successfully to perform opinion mining for languages that do not have enough resources for training specific models.

This study makes the following contributions:

- a robust approach that uses machine translation to reuse a model trained on one language in other languages
- an RNN-based approach to eliminate feature extraction and resource requirements for sentiment analysis
- a technique that outperforms baselines for multilingual sentiment analysis task when data are limited

METHODOLOGY

In order to eliminate the need to find data and build separate models for each language, we propose a multilingual approach in which a single model is built in the language that has the highest number of resources are available. In this paper, we focus on English for building a model because several sentiment analysis data sets are available in English. To make the English sentiment analysis model as generalizable as possible, we train with a data set that has reviews from different domains, such as camera reviews and restaurant reviews. To employ the trained model, test sets are first translated to English via machine translation and then inference takes place. Figure 1 shows our multilingual sentiment analysis approach.

It is important to note that this approach does not use any resource (such as word embeddings, lexicons, or a training set) in any of the languages of the test sets.

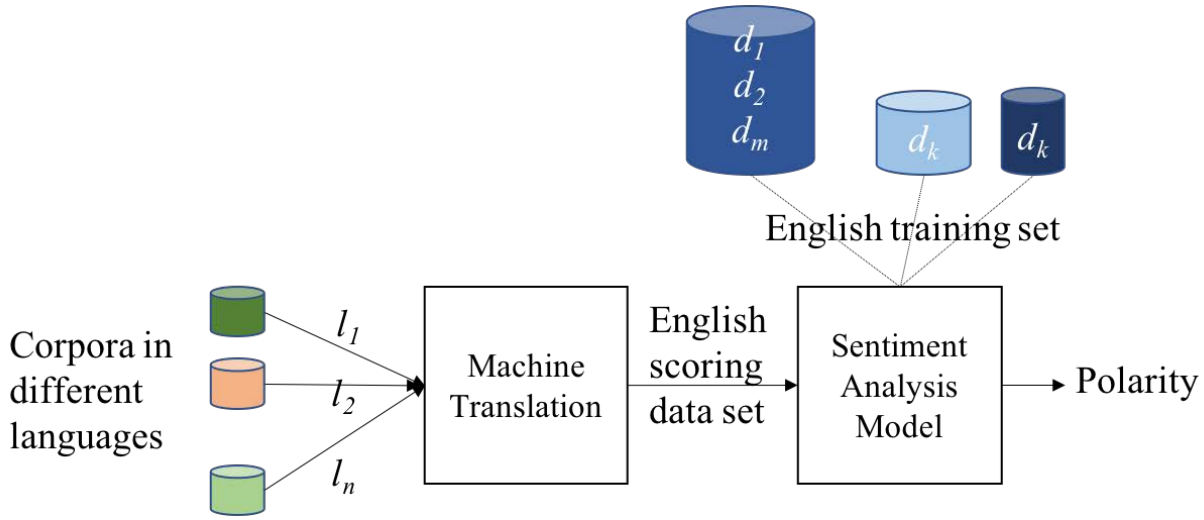

Figure 1. Multilingual Sentiment Analysis Approach

Deep learning approaches have been successful in many applications, ranging from computer vision to natural language processing (Alom et al. 2018). Recurrent neural networks (RNNs) including long short-term memory (LSTM) and gated recurrent units (GRUs) are subsets of deep learning algorithms in which the dependencies between tokens can be used by the model. These models can also be used with variable-length input vectors, making them suitable for text input. LSTM and GRU models allow operations of sequences of vectors over time and have the capability to remember previous information (Alom et al. 2018).

RNNs have been found useful for several natural language processing tasks, including language modeling, text classification, and machine translation. An RNN can also use pretrained word embeddings (numeric vector representations of words that are trained on unlabeled data) without requiring hand-crafted features. Therefore, in this paper, we use an RNN architecture that takes text and pretrained word embeddings as inputs and generates a classification result. Word embeddings represent words as numeric vectors and capture semantic information. They are trained in an unsupervised fashion, making them useful for our task.

The sentiment analysis model that is trained on English reviews has three bidirectional LSTM layers, each with 50 neurons. The training phase takes pretrained word embeddings and reviews in textual format, and then predicts the polarity of the reviews. For this study, an embedding length of 100 is used (that is, each word is represented by a vector of length 100). The training phase is depicted in Figure 2.

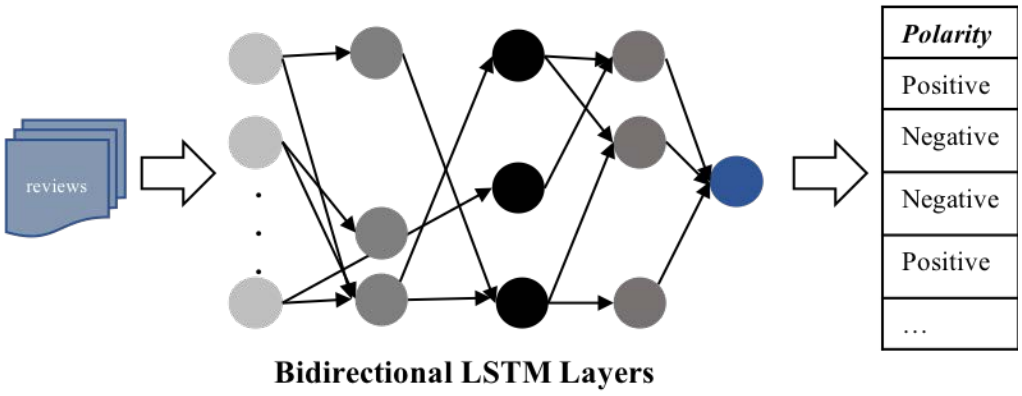

Figure 2. Training Sentiment Analysis Model That Uses LSTM

EXPERIMENTS

To evaluate the proposed approach for our multilingual sentiment analysis task, we conducted experiments. This section first presents the corpora that we used in this study and then shows the experimental results. Throughout our experiments, we used the SAS Deep Learning toolkit.

CORPORA

We used two corpora in this study, both of which are publicly available. The first corpus consists of English reviews, and the second corpus contains restaurant reviews in five different languages (Turkish, Spanish, Russian, Dutch, and Chinese). We focused on polarity detection in reviews; therefore, all data sets in this study have two class values (positive and negative).

For training our LSTM model, we used the data set provided by Kotzias et al. (2015). For evaluation of the multilingual approach, we used five languages. These data sets are part of the SemEval-2016 Challenge Task 5 (Pontiki et al. 2016).

EXPERIMENTAL RESULTS

For experimental results, we reported the majority baseline for each language, where the majority baseline corresponds to a model's accuracy if it always predicts the majority class in the data set. For example, if 60% of all reviews in the data set are positive and 40% are negative, the majority baseline would be 60% because a model that always predicts "positive"' will be 60% accurate and will make mistakes 40% of the time.

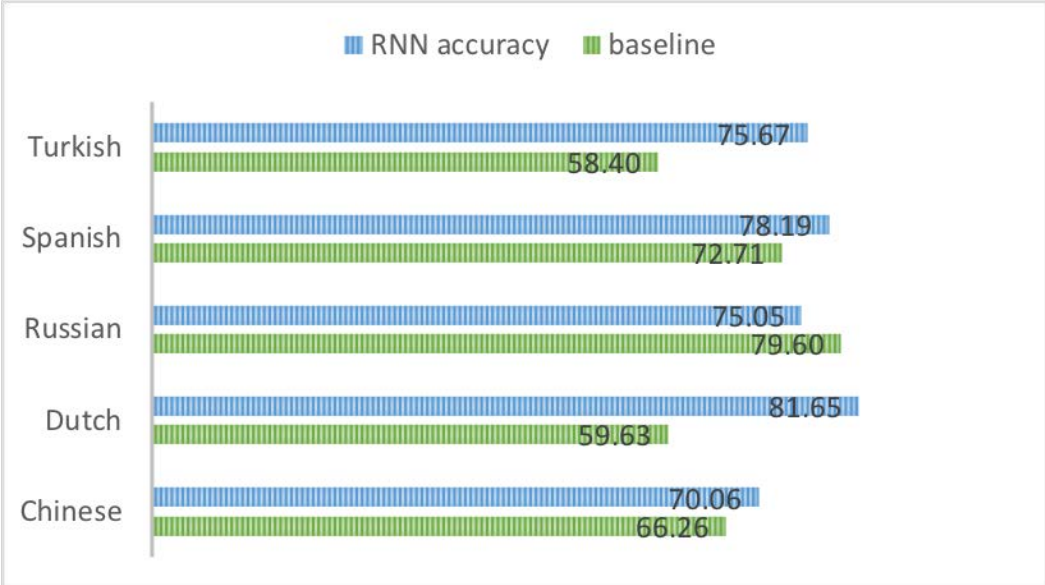

Figure 3. Test Set Accuracy Results for Different Languages

Experimental results on the test set are depicted in Figure 3. An RNN outperforms the baseline in four of the languages (Turkish, Spanish, Dutch, and Chinese). Note that the test set contains reviews from different domains, thus making it challenging for a single model to perform well on all of these languages.

To further analyze how well RNN performed for each language, we calculated the accuracies per class value. This analysis gave us two accuracy values for each language: one for the positive reviews, and one for the negative reviews. Figure 4 shows the accuracies of both positive and negative reviews for all five languages. The general trend is that accuracies of negative reviews are higher than accuracies of positive reviews. This can be explained by the imbalanced nature of the data set: because negative reviews are more frequently found in the data, it is easier for the model to learn from those negative reviews and therefore perform better on that class value.

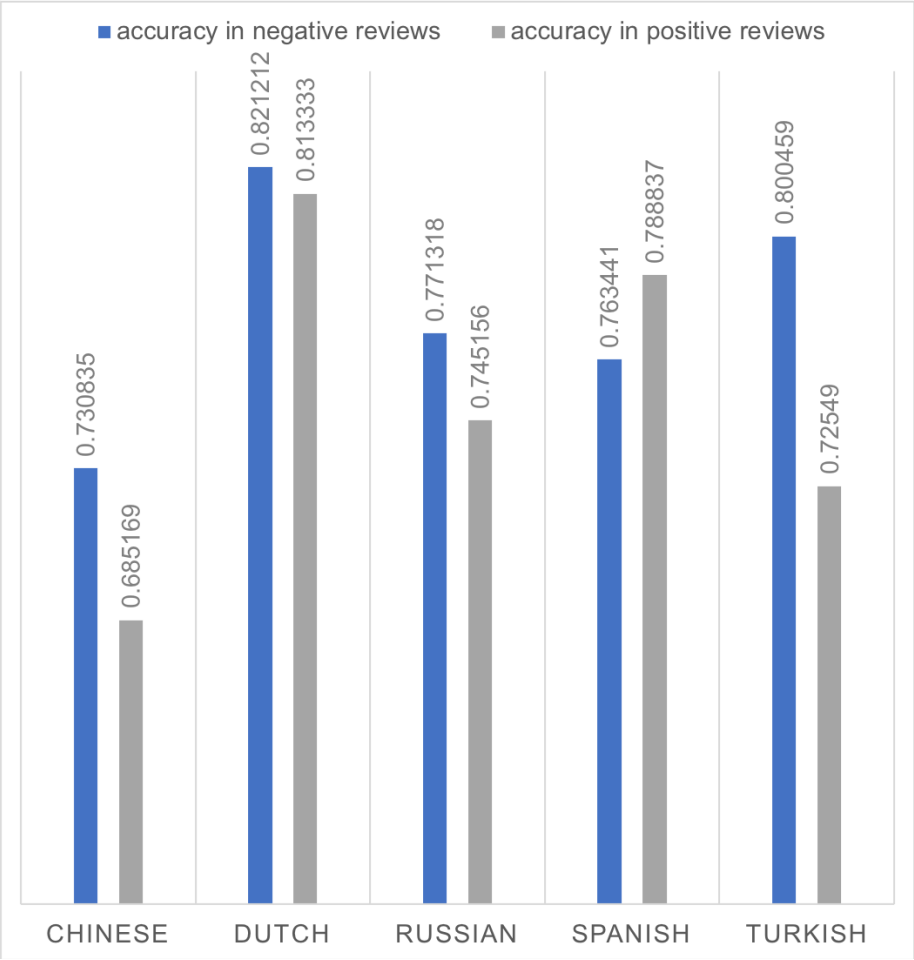

Figure 4. Accuracies of Positive and Negative Reviews

DISCUSSION

One of the crucial elements in using machine translation is to have highly accurate translations. We analyzed the effect of incorrect translations in our approach. To that end, we evaluated our model with an English corpus (Pontiki et al. 2016) to see its performance without any interference from machine translation errors.

Using the English data for testing, the model achieved 79.8% accuracy, where a majority baseline was 68.37%. In the test data, 84% of negative reviews and 77.84% of positive reviews were correctly classified.

Considering the improvements that were achieved by the RNN model over the majority baseline for both English and non-English reviews, we can draw the conclusion that our model is robust in handling multiple languages. Building separate models for each language requires both labeled and unlabeled data. Although having lots of labeled data in every language is the perfect case, it is unrealistic. Therefore, eliminating the resource requirement in this resource-constrained task is crucial. The fact that machine translation can be used in reusing models from different languages is promising for reducing the data requirements.

CONCLUSION

Building effective machine learning models for text requires data and other resources such as pretrained word embeddings and reusable lexicons. Unfortunately, most of these resources are not entirely transferable to different domains, tasks, or languages. Sentiment analysis is one such task that requires additional effort to transfer knowledge between languages.

In this paper, we studied the research question: Can we build reusable sentiment analysis models that can be used for making inferences in different languages without requiring separate models and resources for each language? To that end, we built a recurrent neural network model in the language that had largest amount of data available. During scoring, we used corpora from different domains in different languages and translated them to English to be able to classify sentiments by using the trained model. Experimental results showed that the proposed multilingual approach outperforms the baseline.

REFERENCES

Alom, M. D., T. M. Taha, C. Yakopcic, S. Westbert, P. Sidike, M. S. Nasrin, B. C. Van Esesn, et al. 2018. "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches." arXiv:1803.01164

Banea, C., R. Mihalcea, and J. Wiebe. 2008. "A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources." 2008. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Vol. 8.

Inui, T., and M. Yamamoto. 2011. "Applying Sentiment-Oriented Sentence Filtering to Multilingual Review Classification." *Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP)*. IJCNLP 2011. 51–58.

Kotzias, D., M. Denil, N. de Freitas, and P. Smyth. 2015. "From Group to Individual Labels Using Deep Features." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ortigosa, A., J. M. Martín, and R. M. Carro. 2014. "Sentiment Analysis in Facebook and Its Application to e-Learning." *Computers in Human Behavior* 31: 527–541.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar. M. Al-Smadi, M. Al-Ayoub, et al. 2016. "SemEval-2016 Task 5: Aspect Based Sentiment Analysis." *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Steinberger, J., M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, et al. 2012. "Creating Sentiment Dictionaries via Triangulation." *Decision Support Systems* 53.4: 689–694.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37.2: 267–307.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Ethem Can
SAS
ethem.can@sas.com

Aysu Ezen-Can
SAS
aysu.ezencan@sas.com