Sentiment Classification for Arabic Texts

Abstract

Nowadays, millions of users are expressing their opinions and feelings regarding different life aspects or received products and services, resulting in a massive amount of opinionated texts. Such a rich source had encouraged business, content maker to work with data scientists, diving in texts and obtain useful insights and predictions that enables monitoring comments and feedbacks of their customers or followers and take faster reactions when needed. Sentiment classification deals with the identification of sentiment as either positive or negative. The purpose of this project is twofold. First, building models to discover the sentiment of Arabic sentences as either positive or negative, including testing Arabic pretrained LM and word embeddings on different datasets and compare between them. Second, investigating the impact on performance as a result of using different Arabic text types/domains.

Datasets

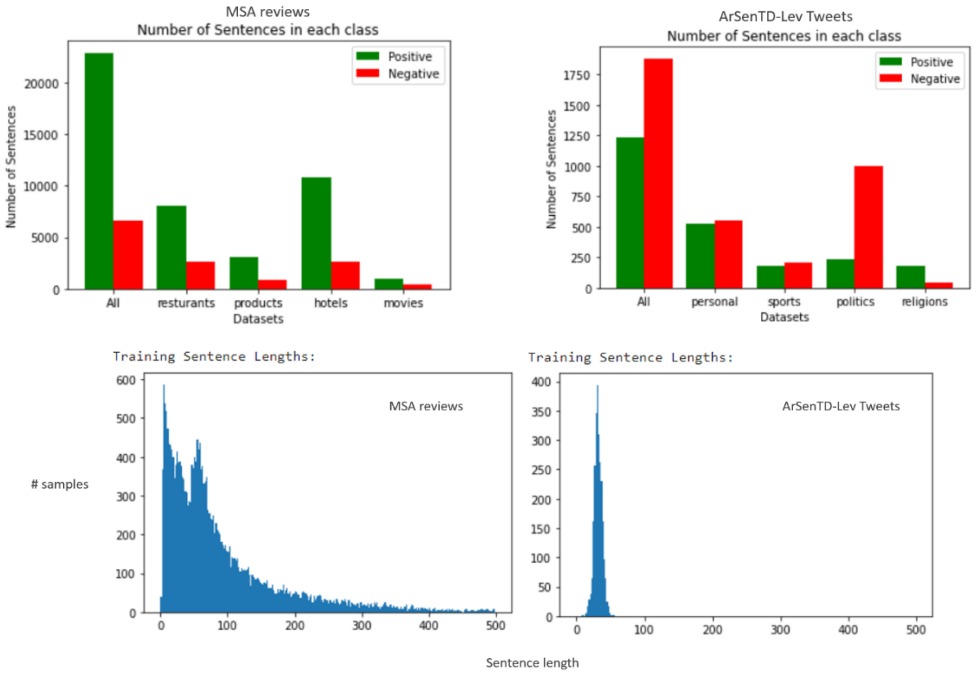
Large Multi-Domain Resources for Arabic Sentiment Analysis

- 29444 reviews in Modern Standard Arabic (MSA)
- Domains: Restaurants, products, hotels, movies.

ArSenTD-Lev (Arabic Sentiment Twitter Dataset for LEVantine dialect)

- 3115 Tweets in dialectal Arabic (Levantine)
- Domains: Personal, sports, politics, religion.

Figures below illustrate class distributions and sentence length distributions on each dataset. More analysis and charts can be found in the notebooks.



Modeling and tools

The following models were trained on Google collab GPU, after classical cleaning and pre-processing for Arabic:

- Linear SVM using TF-IDF on unigrams and bigrams with its default parameters.
- Bi-LSTM using Keras tokenizing (i.e., indexing) and padding.
- Bi-LSTM using Arabic word2vec.
    - A version pretrained on Wikipedia for MSA reviews.
    - A version pretrained on Twitter for ArSenTD-Lev tweets.
- AraBERT tuning (version v02).

Experiments' details can be found in the notebooks.

Results and findings

Dataset	Domain	Size	SVM		AraBert		Bi-LSTM-W2V		Bi-LSTM tokenizing/padding	
			F1	Acc	F1	Acc	F1	Acc	F1	Acc
MSA	Restaurants	10705	0.92	0.87	0.94	0.91	0.90	0.84	0.90	0.85
	Products	3964	0.91	0.86	0.94	0.90	0.91	0.85	0.89	0.82
	Hotels	13422	0.97	0.95	0.99	0.98	0.92	0.87	0.95	0.93
	Movies	1353	0.89	0.83	0.90	0.85	0.84	0.71	0.83	0.72
	All	29444	0.94	0.90	0.96	0.94	0.59	0.49	0.93	0.88
ArSenTD-Lev	personal	1078	0.79	0.79						
	Sports	384	0.82	0.83						
	Politics	1237	0.67	0.90						
	Religions	227	0.92	0.87						
	All	3115	0.76	0.83	0.87	0.90	0.82	0.87	0.75	0.81

F1-score and accuracy results for all experiments are listed in the above table. We can see that AraBERT gave the best results in both MSA and dialects and in all domains. Interestingly, simple linear Support Vector Machine (SVM) overcomes Bi-LSTM in the conducted experiments. However, Bi-LSTM-w2c increased f1-score in ArSenTD-Lev tweets from 0.76 to 0.82. It should be mentioned that increasing Bi-LSTM structure complexity or number of epochs may produce better results, in addition to hyperparameter tuning. Moreover, Bi-LSTM-tokenizing/padding tends to beat or perform better than Bi-LSTM-w2v only for large datasets (i.e., restaurants, hotels and All-MSA).

Conclusions

This project focused on binary sentiment classification on Arabic texts, considering MSA and dialects using reviews and tweets in different domains. There is a substantial room for improvement on the reporting results since they were obtained with no much experiments due to time constraints. Suggestions includes dealing with the variety of sentence lengths (e.g., removing samples with short sentence), handling imbalance (e.g., oversampling or under sampling, and hyperparameter tuning with cross-validation. To conclude, extra efforts to be done to overcome challenges caused by variety of Arabic dialects and its limited resources.