Sentiment Analysis of Ali Babacan Based on eksisozluk.com

Introduction:

This report aims to analyze public sentiment regarding Ali Babacan, a Turkish politician over the course of multiple years. This political figure was chosen due to his recent presence in broadcast and social media and because there exist mixed views towards him. As a data source, the Turkish social media site eksisozluk.com was used and the entries of 26 topics were analyzed. The site eksisozluk.com is a publicly open and accessible site with minimal complexity in terms of data download potential. The site also hosts users of various backgrounds making it a sufficient candidate to be representative of the general population. The site's contents, which are user entries, are made up entirely of text, some of which were labeled by the author as being negative, neutral or positive. Several classifiers were trained based on this labeled data and tested using both count and tfidf vectorized text as input.

Download:

The download process uses the *requests* and *BeautifulSoup* libraries to gather the data and writes it to a csv using the *pandas* library. eksisozluk.com requires calls to be made from legitimate browsers, hence the *user-agent* header has to be modified. The 26 hand-chosen topics yielded over 9,000 entries and the download process takes only about 3 minutes.

Clean-Up:

The entries on the website first need to be cleaned for special characters such as newline etc. as well as for site-specific occasions like references to other entries or websites. Later they need to be checked and corrected for spelling because there are too many words as there is, but also because it increases consistency if all words are spelled correctly. Punctuation marks and spaces were also put in appropriate form. Furthermore, stemming was included as the suffix-additional nature of the Turkish language may lead to too many words, especially nouns, reducing the relational connectedness of the words. It should be noted that stemming does decrease the effect of the specificity of words, and might not have been preferred if the language of the texts were in English, but since they were in Turkish it was decided stemming would be the preferred course.

Labeling:

The data had to be labeled manually into 3 categories, negative, neutral and positive. The sampling was automized to ensure randomness. 750 random entries were selected with an emphasis on the more recent entries as the motivation of the study was the recent media presence of the politician and it was observed that the frequency of entries increased lately. In this way topics and phrases more common would be fully captured, ensuring higher consistency.

Analysis:

The labeled data was used in the training and testing of vectorizers and classifiers. 60% of the labeled data was used as a train set and 40% was used as the test set.

The text input had to be transformed into numerical format, which was done using count and tfidf vectorizers. These vectorizers count the number of occurrences of each word throughout the data and tfidf reduces the weight given to more frequent words. Upon testing with both vectorizers, tfidf was observed to produce results with higher accuracy scores.

Several machine learning classifiers (AdaBoost, RandomForest, SGDC, LogisticRegression, K-Neighbors, SupportVector and more) have been trained and tested multiple times and the average of their accuracy scores were compared. Logistic Regression performed better than the others consistently, with the Random Forest classifier as a runner-up. Due to this performance, Logistic Regression was used through the rest of the study.

Prediction:

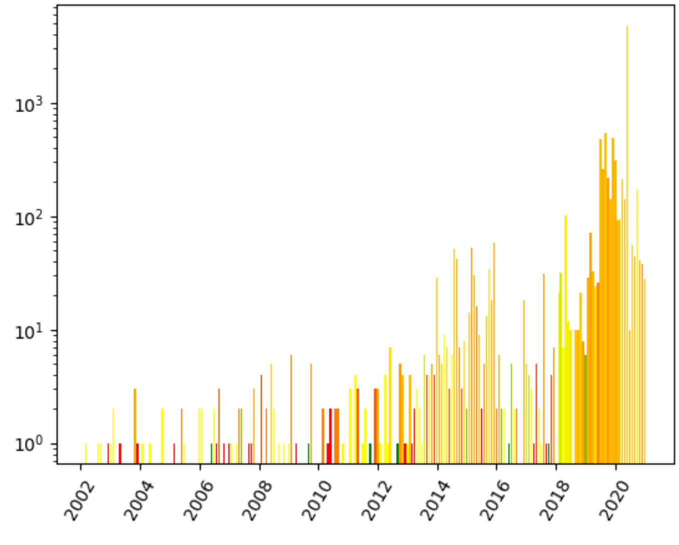
The entirety of the labeled data was used as a train set for the LR classifier and the entire cleaned data was used to predict the sentiment. The results were classified into 3 categories just like the labeled data: into negative, neutral and positive. The tfidf vectorizer had to be limited to accept only words that were also present in the labeled dataset.

Presentation:

The resulting predictions were grouped into monthly data and transformed into a bar chart. The height of the bars represents the number of entries that occurred in that month, whereas the color of the bars were scaled based on the sentiment. As negative was denoted by 0 and positive by 2, the average of sentiments lays between these two bounds. The color red was used to denote negative sentiment versus green for positive sentiment and a linear color scale was used for values in between. The y axis had to be logarithmically scaled as entry frequency increased drastically recently.

Results:

Even the highest accuracy scores produced (using tfidf vectorizer and Logistic Regression classifier) ranged between 0.52 and 0.56. These accuracy levels are significantly below the anticipated 0.8 or even 0.7 levels. The accuracy within the training set reaches 95% levels indicating the classifier is likely working as expected. Based on the classifier's predictions, sentiment regarding Ali Babacan has turned more negative since 2018.



Obstacles:

Data from twitter.com was intended to be included in the study, however APIs provided only data going back one week and other methods of acquiring data (using advanced search and *requests*) was hindered by twitter's extensive security. It was attempted to gather the necessary tokens and cookies using *requests*, and the method is promising as 3 of the 4 required tokens were successfully identified, however the last token was never captured. Using twitter data would have made the dataset more representative of the population as eksisozluk.com may be biased (likely negatively).

Shortcomings:

The stemmer and spell-checker that was used do a fine job but include clearly visible errors. As the data is of Turkish language, more popular libraries could not be used.

A PolynomialFeatures transformation could not be applied as the dataset included over 77 thousand individual words, and even a degree 2 transformation was too large to handle. Due to this fact, meaning originating from word pairs or phrases was lost. There comes no solution to mind on such a large dataset, which is frustrating.

Recommendations:

Translating the data into English before stemming and analyzing may provide better results as the noun structures of English are much more accommodating for sentiment analysis. No such translator library was used in this study.

Special thanks:

Apart from the widely used libraries such as pandas, matplotlib and requests, special thanks are in order for the libraries turkishnlp (which was used for spellchecking) and TurkishStemmer (which was used for stemming). Both libraries were used in the analysis despite their shortcomings and have helped improve accuracy scores.