Classifying Social Media Comments on Brexit Political Stance

Federica Comuni
guscomfe@student.gu.se

Abstract

January 2020 is the month marking the official exit of the United Kingdom from the European Union, usually shortened in "Brexit". Over the years preceding Brexit, Internet users have expressed their support or dissent through comments on social media platforms, forums, and comment sections of online news websites. This study proposes an accuracy comparison of several machine learning classifiers in predicting the pro- or anti-Brexit stance of a corpus of 14680 comments collected from various internet sources, following text pre-processing. The best performing model is a Multinomial Naïve Bayes classifier with an accuracy of 82.2%. The study then proposes an analysis of important features.

Introduction

In June 2016, the United Kingdom held a referendum to decide whether the country should leave or stay in the European Union. The exit option received the majority of votes, thus leading to the so-called Brexit, a gradual withdrawal process culminated with the official exit of the United Kingdom on January 31, 2020 (Barnes 2020). Internet users all over the world, and particularly in the United Kingdom and Europe, have shown their support or disagreement with the decision over the past few years. Social media platforms such as Twitter, Reddit and YouTube, as well as forums and online news websites, have often been sources of comments from users expressing their pro- or anti-Brexit political stance. Meanwhile, natural language processing and machine learning (ML) have been extensively employed to perform classification and sentiment analysis on text data, including social media posts (Kanakaraj and Guddeti 2015). In particular, ML models such as Naïve Bayes and Support Vector Machine have shown promising accuracies (>95%) in text classification tasks, e.g. detecting bullying instances in social media posts (Abdullah-Al-Mamun and Akhter 2018).

This study applies several ML algorithms to the detection of pro- or anti-Brexit political stances in social media posts and comments, collected on various websites in 2019 and 2020 by students of the course Applied Machine Learning at the University of Gothenburg and Chalmers Institute of Technology. The compared classifiers are a Linear Support Vector classifier, a Multinomial Naïve Bayes classifier, different types of linear classifiers with Stochastic Gradient Descent (SGD) and a Logistic Regression classifier. An analysis of the most important features is also proposed. The following section describes in more detail the pre-processing of the data and the classification process.

Methods

Data collection

The data were collected in a joint effort by the students of the Applied Machine Learning course, from several online sources. Each comment was labeled with pro- (1) or anti- (0) Brexit stance by the student who first collected it. In a following round, students were given the data collected and labeled by other students and they were asked to assign an additional label to each comment blindly, i.e. without seeing the original annotator's label. The students had the option to mark the comment with a -1 if they did not understand its stance. At the end of this phase, each comment was assigned between two and six labels, depending on how many students had reviewed it. The data set was then split in 92% training set (13520 comments) and 8% test set (1160 comments).

Text pre-processing

A few rounds of text pre-processing were performed to remove noise from the text and allow the machine learning models to work on the most salient features. First, the texts were stripped of punctuation marks except for the exclamation point, the question mark, and the hash symbol. The latter was left to recognize social media hashtags, while the first remained because upon inspection of the data they seemed preponderant in pro-Brexit comments, and were later found to improve the accuracy of the models. Secondly, comments were converted to lower case and cleaned from the standard Natural Language Toolkit (NLTK) set of English stop words, excluding "against", "no", "not", "don" and "don't", which were left in the text because considered indicative of political stance. Hashtags were then identified by the presence of the hash symbol, and comments with hashtags were segmented using the wordsegment library to split hashtags into single words. At last, words were lemmatized and stemmed. A TfidfVectorizer was then fit to the data while set to further remove accents and to create features from word unigrams, bigrams, trigrams and quadrigrams.

Additional features

Two additional features were added to the input data from the vectorized word matrix: the comments' length, and the probability of each comment in the training set of belonging to one of the two topics found through Latent Dirichlet Allocation (LDA) topic modeling. For the first feature, the length in characters of each sentence after pre-processing was considered. For the second feature, a bag-of-words corpus of the words in the training set was used to find the two most frequent topics; the probability of each comment belonging to the topics was then calculated thanks to the corpus. Both feature vectors were then added to the vectorized word matrix to be used as input for the ML models.

Classifiers

The classifiers were implemented using the Scikit-Learn library. Grid search 5-fold cross validation was employed to find the optimal parameters for each classifier and to train and evaluate the models. The classifier algorithms were chosen among those that performed with the highest accuracy, compared to the 51.1% accuracy of the dummy classifier. For the training data, the study took into consideration the different annotators' labels by discarding the -1 labels and computing the mean of the remaining labels. The original annotators were assumed to have greater insight on the political stance of the comment, therefore the original label weighed more on the mean (1.3 times more than the other labels). The final label was then set to 1 if the mean was equal to or greater than 0.5, 0 otherwise.

Results

The models classified the test set with accuracy between 79.2% for Logistic Regression and 82.2% for the Multinomial Naïve Bayes. Figure 1 compares the accuracy, precision and f1 score of the four models. The Multinomial Naïve Bayes also presented the highest f1 score.

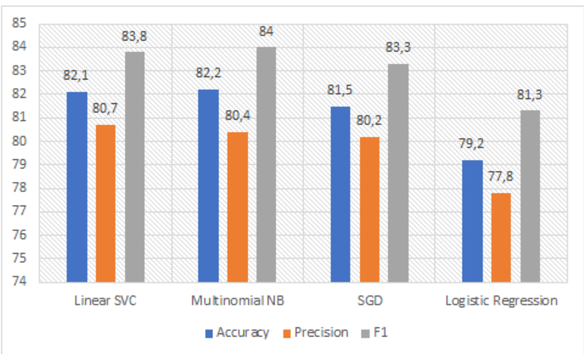

Figure 1: Accuracy, precision and f1 score of the models

Table 1 shows a few of the optimal parameters for each classifier found through grid search. The accuracy rates shown in figure 1 were achieved after discarding the length and topic features, which yielded a slightly lower accuracy both on the validation set and test set (0.6% lower accuracy on the validation set, 0.4% on the test set, for the Multinomial Naiv̈e Bayes classifier). Excluding "?" and "!" from removed punctuation also yielded higher accuracy (0.7% more accuracy on the test set for the Multinomial Naïve Bayes). The lower accuracy for the training set with topic modeling might have been due to a lack of suggestiveness of political

Linear SVC	Multinomial NB
loss = "squared hinge"	alpha = 1.0
penalty = "l2"	class_prior = None
max_iter = 1000	fit_prior = False
SGD	Logistic Regression
loss = "squared hinge"	penalty = "l2"
early_stopping = False	
max_iter = 1000	

Table 1: Optimal parameters for each model

stance in the topics, or to a low distinctiveness between the two topics. Table 2 shows the 10 most common lemmas in the two topics: the words in the first topic might belong more often to an anti-Brexit stance and the words in the second to a pro-Brexit, but this categorization is only hypothetical and the distinction does not seem to be clear enough for this feature to be relevant.

Topic 1	Topic 2
EU	leave
UK	EU
Britain	Brexit
country	vote
Europe	people
good	not
not	no
european	want
go	deal
great	UK

Table 2: Most frequent lemmas in topics

Figure 2 shows the models' sensitivity and specificity. For all models, the sensitivity was considerably higher than specificity, meaning that positive entries (pro-Brexit) were classified more accurately than negative entries (anti-Brexit). This disparity between sensitivity and specificity (as high as 15% for the Logistic Regression model) might be explained by the consistent presence in the data set of neutral comments that were originally annotated as anti-Brexit. Also, it might be explained by the presence of comments expressing sarcastic enthusiasm towards Brexit.

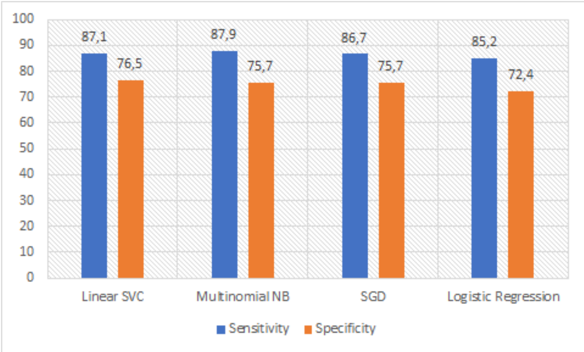

Figure 2: Sensitivity and specificity of the models

Upon inspection of the annotations from different students in the training data, it appears that there is a considerable amount of instances in which the annotators disagreed, around 1 in 10 comments. In a few cases, it appears that the disagreement might have been due to the comment requiring some degree of detailed knowledge about Brexit that one of the annotators lacked (for example, one comment insulted ”46.6% of voters”, probably referring to the 46.6% of the English population who voted to remain). Other disagreements concerned sarcastic comments or comments talking about something not directly related to Brexit. Analyzing examples of misclassified test set comments can shed some light on the reason for the misclassification shown in figure 2. The first example of misclassified comment (a false positive) seems to be a reply to a previous comment, and it is not immediately clear whether it is pro- or anti-Brexit:

British people are p'wned by banks.

A second example (this time a false negative), similarly to the first one, does not appear very clear:

We had our vote forget about a second!!!

A third example (a false negative), although very clear to a human reader, shows vocabulary from both stances, thus confusing the classifier:

I've changed my mind from remain to leave.

A fourth example (false positive) includes sarcasm:

Farewell Great Britain send our regards to the Earth's core on your free fall into oblivion

After looking at the misclassified examples, it is reasonable to believe that most misclassifications occurred because of neutrality or sarcasm in comments, and because the comments were not directly pertinent to Brexit.

At last, figure 3 shows the 15 most important text features, found by an additional Random Forest classifier trained for the purpose. The three most relevant words are ”EU”, ”Brexit” and ”leave”, as it can be seen in the labels in figure 4. 7 words out of 15 coincide with the words found by LDA topic modeling, including the three most relevant, as previously shown in table 2.

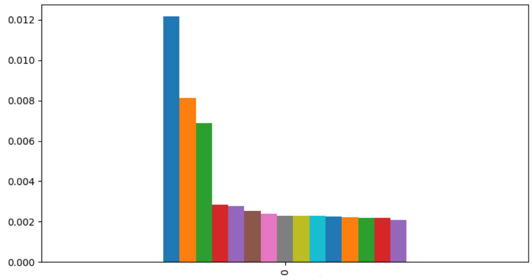

Figure 3: Most important text features

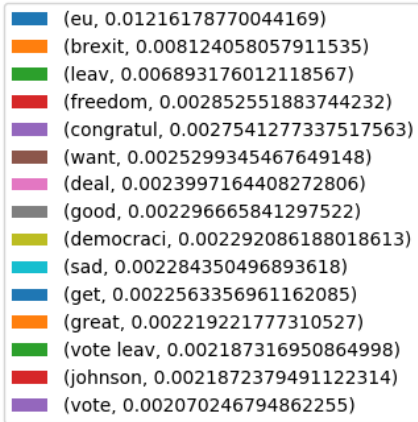

Figure 4: Feature names and importance

Conclusion

The study compared the accuracy of several ML classifiers in detecting pro- or anti-Brexit stance in social media comments. After a few rounds of text pre-processing and hyperparameters tuning, the classifiers were able to reach accuracy rates of 81.25% on average, with a Multinomial Naïve Bayes classifying with 82.2% accuracy as the best model. Analysis of the classification of the training data and misclassified test set examples revealed some uncertainty and error both in human and machine classifiers, possibly due to sarcasm or neutrality in the comments. Accuracy might benefit from using more complex models such as deep neural networks, which might be capable of detecting sarcasm and might be more sensitive to small nuances in neutral comments than non-neural ML models.

References

Abdullah-Al-Mamun, and Akhter, S. 2018. Social media bullying detection using machine learning on Bangla text. *2018 10th International Conference on Electrical and Computer Engineering (ICECE).*

Barnes, P. 2020. Brexit: What happens now? *BBC News.*

Kanakaraj, M., and Guddeti, R. M. R. 2015. NLP based sentiment analysis on Twitter data using ensemble classifiers. *2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN).*