FinalReport

September 27, 2023

1 Sentiment Analysis on Machine Translated Icelandic corpus

- Ólafur Aron Jóhannsson
- Eysteinn Örn
- Birkir Arndal

2 Contents

2.1 Abstract

Translating English text into low-resource languages and assessing sentiment is a subject that has received extensive research attention for numerous languages, yet Icelandic remains relatively unexplored in this context. We leverage a range of baseline classifiers and deep learning models to investigate whether sentiment can be effectively conveyed across languages, even when employing machine translation services such as Google Translate and Miðeind machine translation.

2.2 Introduction

In this research endeavor, we utilized an IMDB dataset comprising 50,000 reviews, each categorized as either positive or negative in sentiment. Our methodology involved the translation of these reviews using both Google Translate and Miðeind Translate. Subsequently, we subjected all three datasets, including the original English version and the two translations, to analysis using three baseline classifiers. The primary objective was to investigate whether machine translation exerted any influence on the results of sentiment analysis and to determine the superior performer between Miðeind and Google translations. Our aim was to assess the transferability of sentiment across machine translation processes.

2.3 Machine Translations

We employed the Google Translator API, which relies on Google's Neural Machine Translation featuring an LSTM architecture. Additionally, we utilized the Miðeind Vélþýðing API for the purpose of machine-translating the reviews. The Miðeind Vélþýðing API is constructed using the multilingual BART model, which was trained using the Fairseq sequence modeling toolkit within the PyTorch framework.

2.3.1 Google Translate

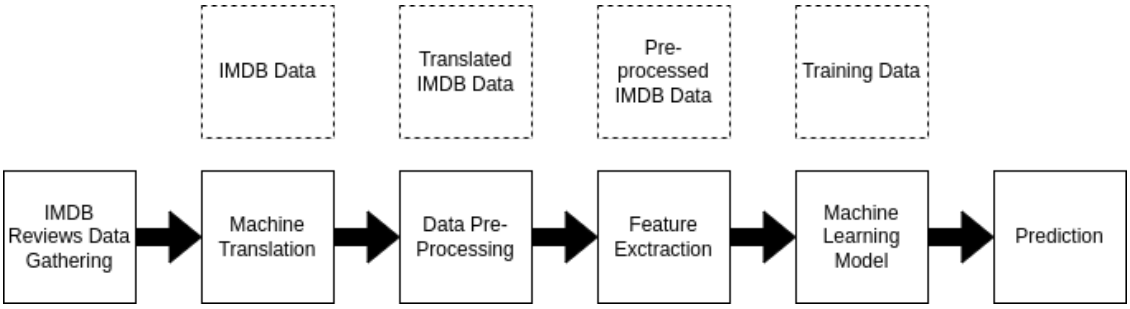
All the reviews were effectively translated using the API, and the only preprocessing step performed on the raw data was the removal of
. The absence of errors during the translation process could be attributed to the API's maturity and extensive user adoption. Nevertheless, it's worth noting that the quality of Icelandic language reviews occasionally exhibited idiosyncrasies.

2.3.2 Miðeind

The Miðeind Translator encountered challenges when translating the English corpus into Icelandic. To prepare the text for translation, several preprocessing steps were necessary. These steps included consolidating consecutive punctuation marks, eliminating all HTML tags, ensuring there was a whitespace character following punctuation marks, and removing asterisks. Subsequently, we divided the reviews into segments of 128 tokens, which were then processed in batches by the Miðeind translator.

2.4 Pre-Processing and feature extraction

The original English dataset we lowercased, tokenized and lemmatized and removed stop words, we applied the same on the Icelandic machine translated corpus as well.



3 Baseline Classifiers

We created pipelines for the three classifiers which serve as a baseline metric for our scoring for English and machine translated Google and Miðeind datasets, all classifiers use TF-IDF vectorizer, which measure the frequency of a term in each document. It measure how important the term is across all documents.

Here is an example of

3.0.1 Original English Dataset

3.0.2 Google Translate Icelandic Dataset

Classifier	Precision	Recall	F1-Score
MultinomialNB			
negative	0.8448	0.8810	0.8625
positive	0.8770	0.8398	0.8580
SVC			
negative	0.8977	0.8907	0.8942
positive	0.8926	0.8995	0.8960
Logistic Regression			
negative	0.8963	0.8808	0.8885
positive	0.8840	0.8991	0.8915

3.0.3 Miðeind Icelandic Dataset

3.1 Naive Bayes

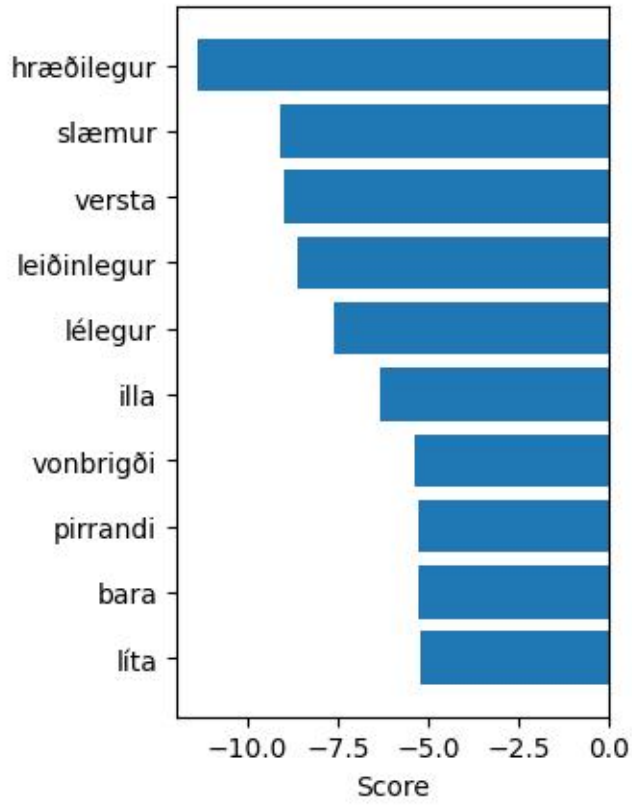
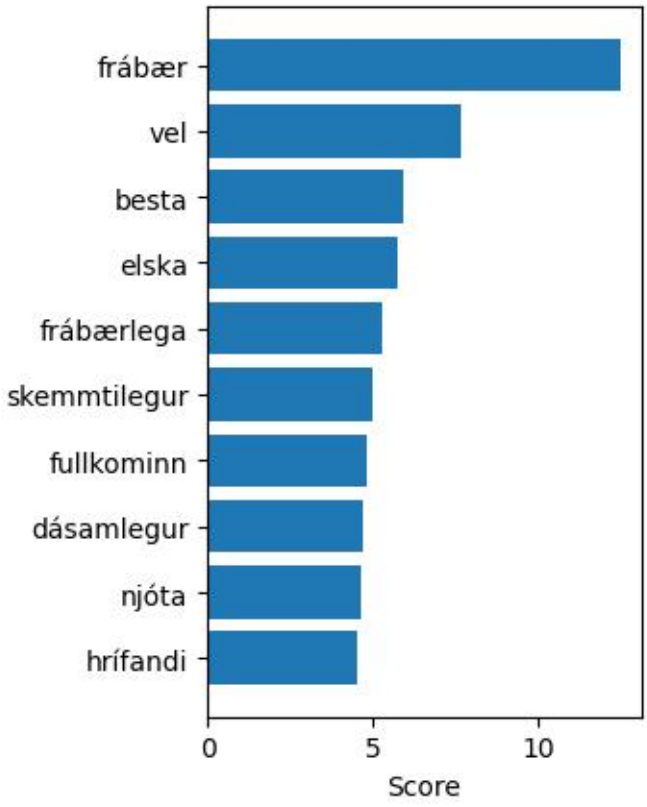
3.2 Logistic Regression

Logistic Regression is a binary classification algorithm, were the result is defined as zero or one in binary models. When we trained the class it gives us a list of coefficients that represent the relationship between the input variables and the output variable in the model. The coefficient can be interpreted as the relative importance of the word it's classified to, in this case negative or positive.

In this chart we can see the top 10 negative and positive values, for a sentence to be positive in this case, it has to have a value of one.

Some examples are after running tests

- (hræðilegur frábær) Positive, score is 1.124940
- (slæmur vel besta) Positive, score is 4.491666
- (lélegur vel) Negative, score is 0.107679

Positive	Negative
	

3.3 Support Vector Machines

4 Models

5 Results

6 Conclusions