FinalReport

September 30, 2023

1 Sentiment Analysis on Machine Translated Icelandic corpus

- Ólafur Aron Jóhannsson
- Eysteinn Örn
- Birkir Arndal

Leiðbeinendur - Hrafn Loftsson (hrafn@ru.is) - Stefán Ólafsson (stefanola@ru.is)

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, with 25.000 being positive and 25.000 being negative. 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.

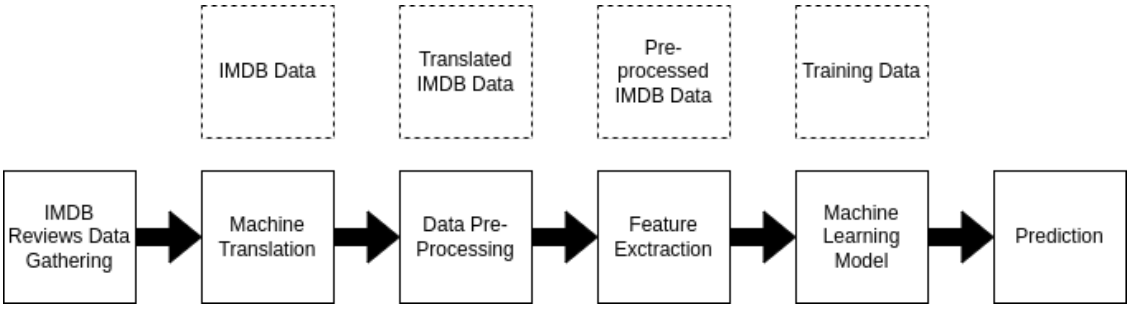
2.3.2 Miðeind

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, the same was applied on the Icelandic machine translated corpus as well, in addition we also added a prefix _NEG to the words in Icelandic if the term was deemed negative to assist the vectorizer in locating negative remarks.

Three baseline classifier pipelines were created that 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. We see scoring of these terms in (#logistic)



3 Baseline Classifier Evaluation

We utilized the classifiers available in the Scikit-learn Python package for implementing our machine learning models. These models were trained with their default parameters, and hyperparameter

tuning was not conducted. It is important to note that superior results can be attained by fine-tuning the hyperparameters.

When assessing the statistical measures to gauge the model's performance, we applied equations 1, 2, 3, and 4.

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{1}$$

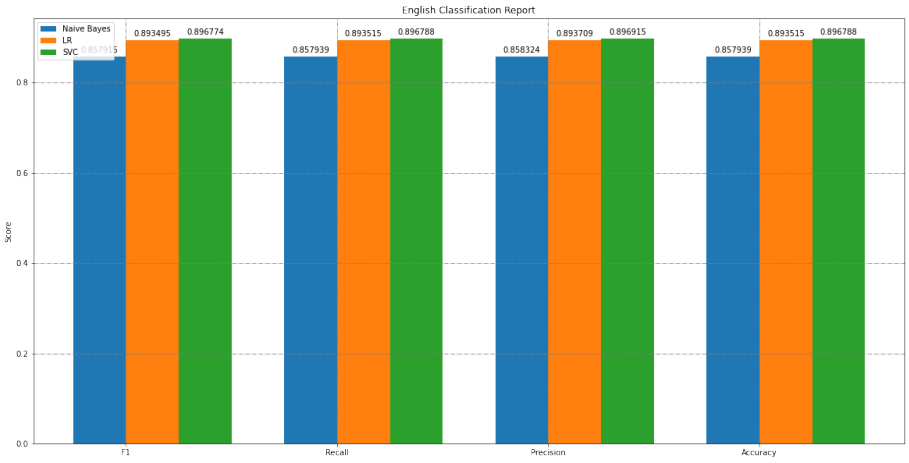
$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

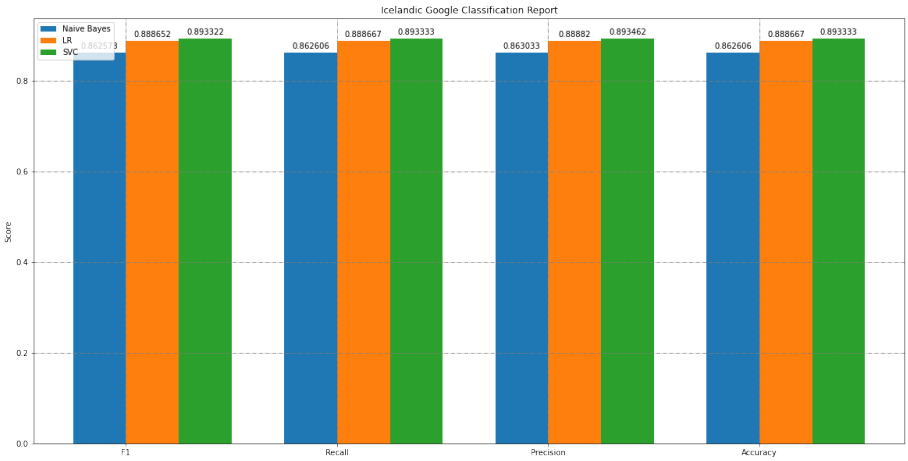
$$F1Score = \frac{2(Recall * Precision)}{Recall + Precision} \tag{4}$$

True Positive (TP) refers to correctly identified positive sentiments, while False Positive (FP) signifies incorrectly identified positive sentiments. True Negative (TN) denotes correctly identified negative sentiments, and False Negative (FN) represents incorrectly identified negative sentiments.

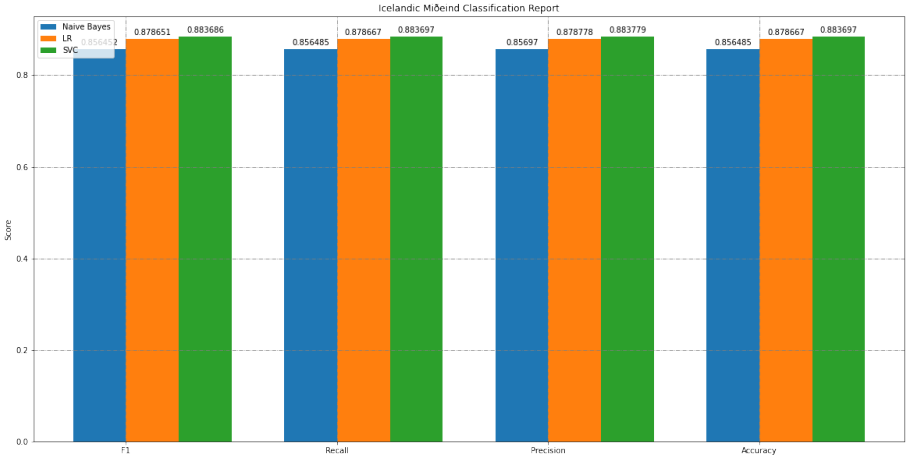
The data was divided into training and test sets, with 67% (33,500 reviews) allocated for training the models and 33% (16,500 reviews) reserved for testing the model's performance.



#



#



#

In this visual representation of the classification report encompassing all classifiers, we observe that Support Vector Classification (SVC) outperformed other models when applied to the data. All models were trained with 33,500 reviews and tested with 16,500. If we establish SVC as our baseline comparative model and employing a weighted F1 score as our evaluation metric, we can discern the following results across different datasets: In the English dataset, the F1 score reached 89.67%, the translated Miðeind dataset achieved an F1 score of 88.36%, and the Google dataset attained an F1 score of 89.33%. These figures suggest that sentiment analysis can carry across Machine Translation when utilizing state-of-the-art machine translation APIs. The loss in accuracy during translation is minimal, with only a 1.31% and 0.34% drop in accuracy, favoring Google's performance.

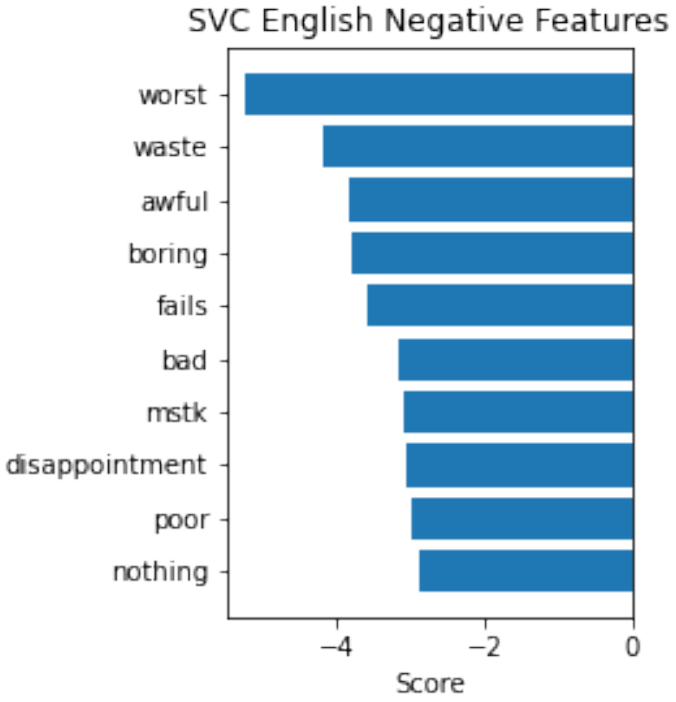
3.1 Support Vector Classifier

The SVC (Support Vector Classifier) was the best machine learning algorithm in classifying sentiment, it is a linear 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.

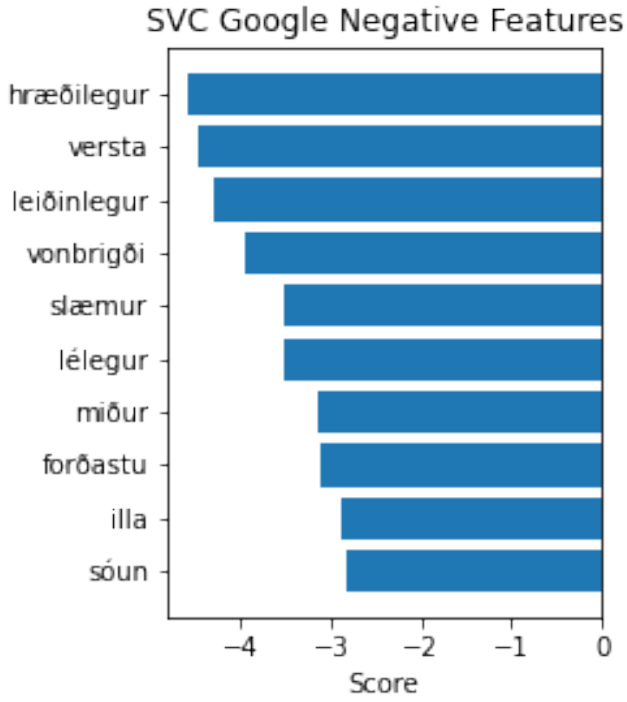
English Sentiment	Precision	Recall	F1-Score
negative	0.90	0.89	0.90
positive	0.89	0.91	0.90

Negative

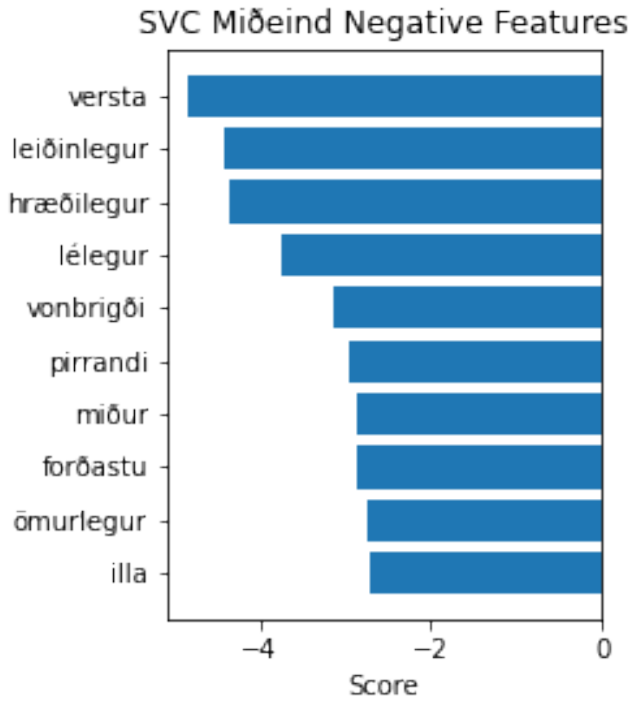


Google Sentiment	Precision	Recall	F1-Score
negative	0.	0.	0.
positive	0.	0.	0.

Negative



Miðeind Sentiment	Precision	Recall	F1-Score
negative	0.	0.	0.
positive	0.	0.	0.

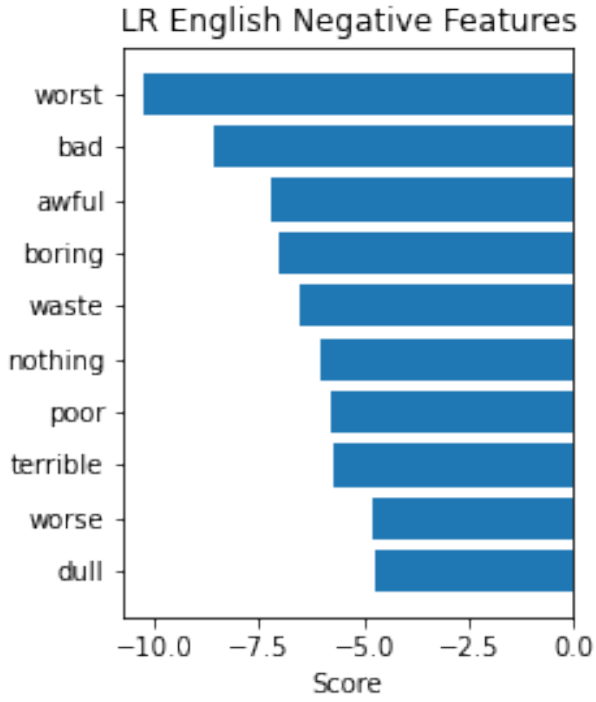
Negative


3.2 Logistic Regression

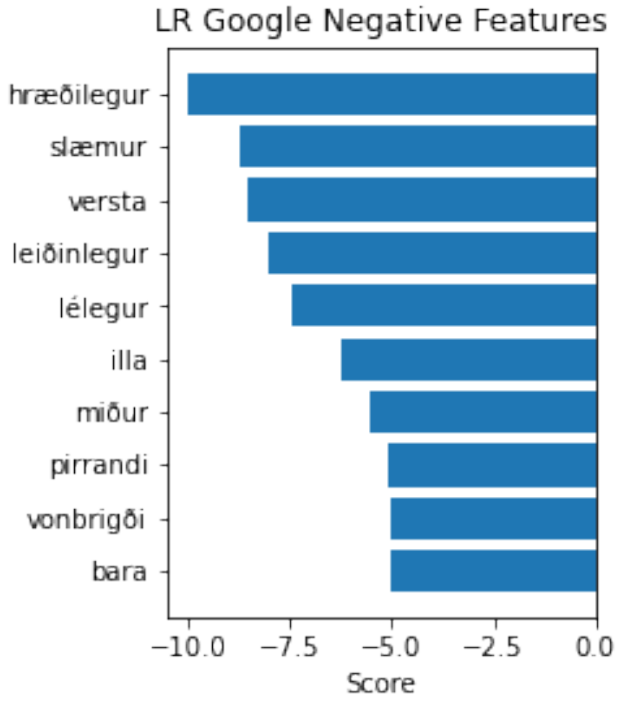
Logistic Regression is a binary classification algorithm, were the result is defined as zero or one in binary models. When we trained the class it gives us a list of coefficients that represent the relationship between the input variables and the output variable in the model. The coefficient can be interpreted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

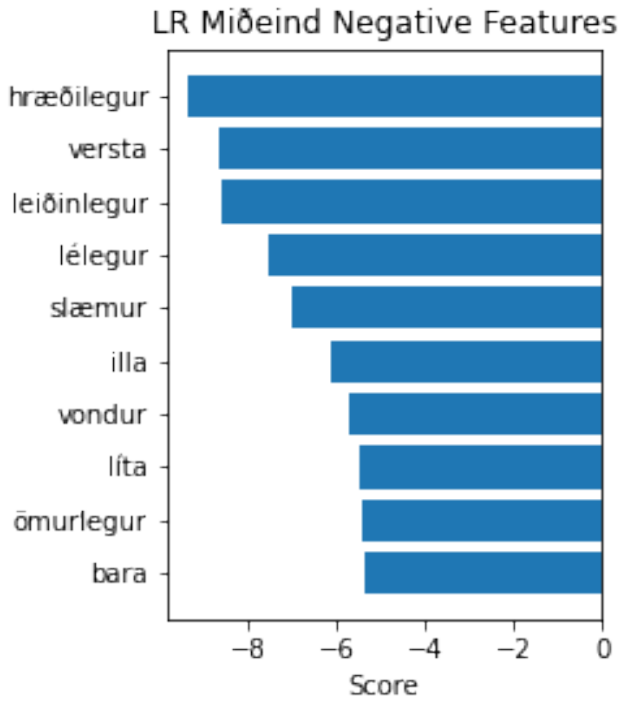
English Sentiment	Precision	Recall	F1-Score
negative	0.90	0.88	0.89
positive	0.89	0.91	0.90

Negative


LR English Negative Features
worst
bad
awful
boring
waste
nothing
poor
terrible
worse
dull
−10.0
−7.5
−5.0
−2.5
0.0
Score

Negative


LR Google Negative Features
hræðilegur
slæmur
versta
leiðinlegur
lélegur
illa
miður
pirrandi
vonbrigði
bara
−10.0
−7.5
−5.0
−2.5
0.0
Score

Negative


LR Miðeind Negative Features
hræðilegur
versta
leiðinlegur
lélegur
slæmur
illa
vondur
líta
ömurlegur
bara
−8
−6
−4
−2
0
Score