FinalReport

November 3, 2023

1 Sentiment Analysis on Icelandic text using Neural Networks and Machine Learning Classifiers

Nemendur - Ólafur Aron Jóhannsson, Eysteinn Örn, Birkir Arndal

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

1.1 Abstract

Sentiment Analysis involves the process of categorizing the general sentiment expressed within a text regarding its subject.

In this research paper, we delve into an experiment focused on sentiment analysis of Icelandic text. Our approach involved training machine learning classifiers and transformer models using machine-translated Icelandic. The primary goal of our investigation was to evaluate the effectiveness of a sentiment classification system when applied to Icelandic text.

Three machine learning classification techniques were employed: Logistic Regression, Naive Bayes and Support Vector Machines which served as a baseline for our research.

Additionally, we performed downstream training of pre-trained transformer models on the machine-translated Icelandic text, utilizing three different iterations of the RoBERTa-based transformer model.

1.2 Previous Work

The topic of machine translating English text into low-resource corpus and examining what impact is has on sentiment classification has garnered considerate research focus for various low-resource languages [1] [3] [4] [5]. However, Icelandic remains relatively underexplored within this domain [2].

1.3 Introduction

Our motivation for this research endeavour is that there is no Icelandic dataset tailored for sentiment analysis that is open and readily accessible to everyone, and creating one from scratch can be an expensive process, especially for low-resource languages. Utilising machine translation serves as an inexpensive method to create such a dataset and allows us to explore whether similar sentiment analysis results can be emulated.

Sentiment analysis involves natural language processing and text analysis to identify, extract, and quantify subjective information, such as positive, negative, or neutral sentiments. It has utility for many applications, such as gauging public opinions, enabling businesses to ascertain and categorise customer satisfaction, and providing valuable insights into user-generated content across diverse digital platforms, e.g., customer reviews, complaints, and comments.

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.

We further assessed the performance of our classifiers and deep learning models using written reviews sourced from an Icelandic website, which comprised 932 positive reviews and 179 negative reviews.

Our hypotheses are as follows:

- Sentiment classification on English text will yield the most favorable outcomes when trained on RoBERTa-base.
- Transformer models and machine learning classifiers trained using machine translations generated by Google Translate will achieve the highest accuracy.
- Sentiment classification on Icelandic text will produce the most optimal results when trained on IceBERT in conjunction with Google Translate.

1.4 Metrics and Evaluation

When assessing the statistical measures to gauge the model's performance, we used an F1 score of each class (positive, negative) - how to calculate the F1 score can be seen in 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.

1.5 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.

1.5.1 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 <br/>. 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.

1.5.2 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 Machine Learning Classifiers

We utilized Support Vector Machines, Logistic Regression and Naive Bayes which are 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.

2.1 Pre-Processing, feature extraction and tokenization

Pre-processing is the act of transforming raw data to a form that can be used for the next part of the machine learning process.

Here are the preprocessing steps that we used:

- Data cleaning: Removing errors and irrelevant data.
- Tokenization: Breaks sentences or paragraphs into individual words.
- Lower casing: Helps normalize and reduce dimensionality of the dataset.
- Lemmatization: Convert different forms of the same word to a standardized form. This reduces the number of unique words in the dataset, which helps to improve the performance of NLP tasks.
- Mark negation: the portion of text that follows a negation word and up to a punctuation, will be marked with _NEG suffix. The purpose of marking negation is to help NLP models to understand the context of a sentence.

We do preprocessing to get a reliable dataset for the machine learning algorithms to give us good performance and an accurate result.

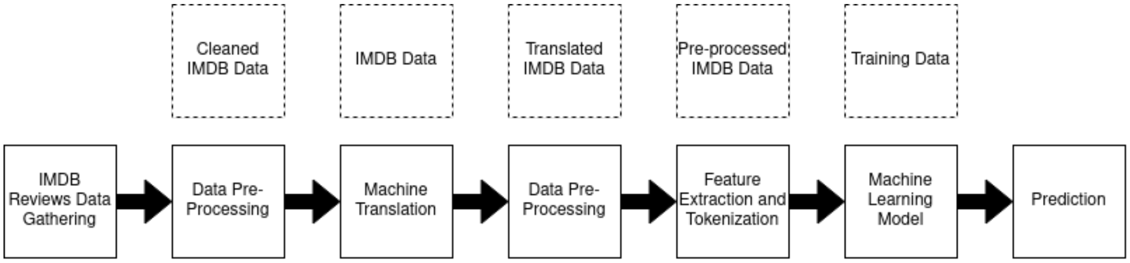
For the original English dataset, we removed noise (e.g., removed html tags), lowercased, tokenized, lemmatized, removed stop words and added a _NEG prefix when a word was starting with negation to assist the vectorizer locating negative sentiments.

For all the Icelandic datasets we removed noise, punctuations except for abbreviations, stop words. We also lower cased all the texts, tokenized, lemmatized and marked negations in the texts.

For the Miðeind translated dataset we additionally had to remove long nonsense words (e.g., "...BARNABARNABARNAÞÁTTURINN"), and replace repeated character to one (e.g., "jááááááá" -> "já").

For the Hannes Official Station dataset we had to do conversion since we got the data as a MySQL file, so we had to escape unicode decimal points for Icelandic characters, remove HTML, special characters, lowercase, remove other noise such as smileys and brackets.

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 and also tokenizes the word for the classifiers. It measures how important the term is across all documents. We see scoring of these terms in (3.2)



2.2 Training

2.2.1 IMDB Movie Review Data

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 classifiers's performance.

In the processed Icelandic machine translated data and the English data we observed that Support Vector Classification (SVC) outperformed other classifiers when applied to the data. Although the F1-score was only fractionally better then Logistic Regression, where the F1 score for the English dataset was 89.53% and 89.82%, for the Machine translated Icelandic Google dataset it was 88.91% and 89.13% and for Miðeind it was 88.00% and 88.29% for negative and positive respectively.

2.2.2 Hand written Icelandic Data

We also obtained a dataset comprising 1,111 movie reviews provided by Hannes [6], the operator of officialstation.com, who specializes in Icelandic movie reviews. These reviews were used to further assess the performance of our classifiers. Hannes assigns numerical scores to the reviews, and we modified the scoring system to categorize reviews with scores between 1 and 5 as negative and those with scores between 6 and 10 as positive.

Our observations revealed that in this scenario, Logistic Regression trained with Miðeind Vélþýðing exhibited the most favorable performance, achieving an F1 score of 45.36% for negative sentiment and 89.22% for positive sentiment. This suggests that the classifiers excel in predicting positive sentiment but encounter challenges in identifying negative sentiment. This challenge could be attributed to an insufficient amount of negative data or the specific wording of the negative reviews not being overtly "negative" enough for the classifier to categorize them accurately.

It's also worth noting that the hand-written Icelandic reviews use much more slang and Icelandic-English words that are in-between, which are not seen in the IMDB training dataset, this is peculiarity in the language because slang is common and the joining of Icelandic/English words into a words that is a hybrid of the two languages, and also having lots of English slang in between.

2.3 Machine Learning Prediction Results

English Sentiment	Precision	Recall	F1-Score
SVC negative	90.38	88.69	**89.53**
SVC positive	89.01	90.65	**89.82**
LR negative	90.22	88.15	89.17
LR positive	88.53	90.55	89.52
NB negative	84.62	87.32	85.95
NB positive	87.04	84.29	85.64

Google Sentiment	Precision	Recall	F1-Score
SVC negative	89.60	88.24	**88.91**
SVC positive	88.47	89.81	**89.13**
LR negative	89.58	87.62	88.59
LR positive	87.94	89.86	88.89
NB negative	84.77	87.14	85.94
NB positive	86.84	84.43	85.62

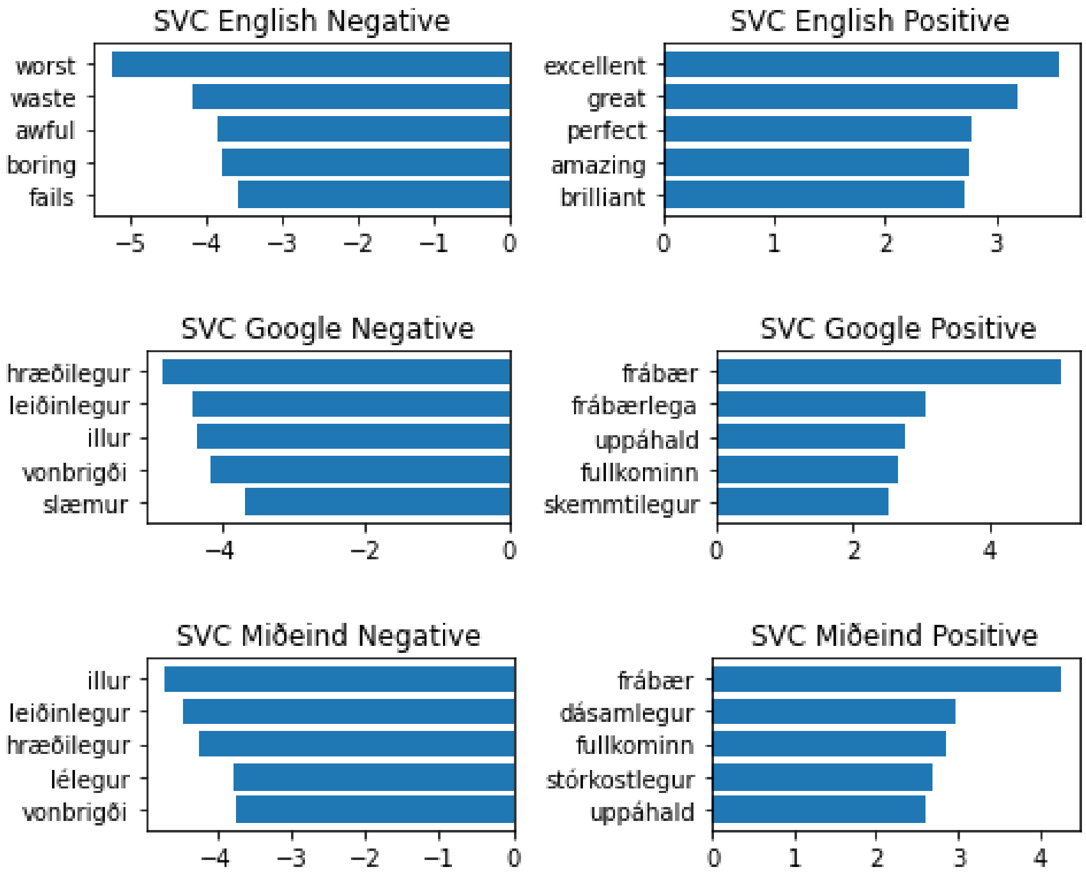
Miðeind Sentiment	Precision	Recall	F1-Score
SVC negative	88.64	87.37	**88.00**
SVC positive	87.67	88.92	**88.29**
LR negative	88.56	86.57	87.56
LR positive	87.00	88.93	87.95
NB negative	84.29	86.23	85.25
NB positive	86.05	84.09	85.06

Hannes Google Sentiment	Precision	Recall	F1-Score
SVC negative	37.86	43.58	40.52
SVC positive	88.84	86.27	87.53
LR negative	42.13	46.37	**44.15**
LR positive	89.50	87.77	**88.62**
NB negative	24.01	60.89	34.44
NB positive	89.35	62.98	73.88

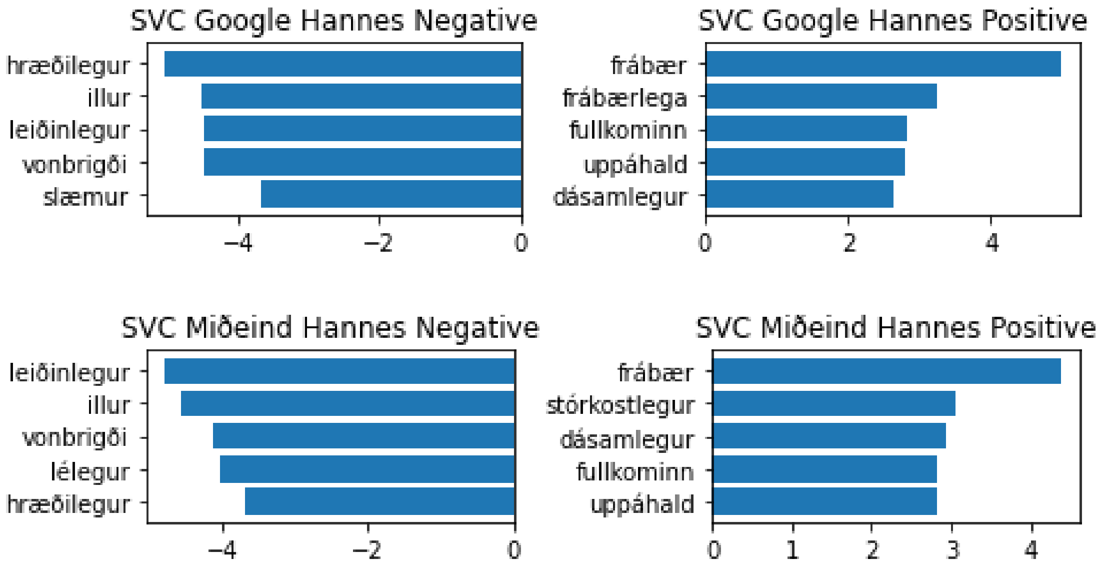
Hannes Miðeind Sentiment	Precision	Recall	F1-Score
SVC negative	38.46	44.69	41.34
SVC positive	89.04	86.27	87.63
LR negative	44.39	46.37	**45.36**
LR positive	89.61	88.84	**89.22**
NB negative	25.50	56.98	35.23
NB positive	89.17	68.03	77.18

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 5 negative and positive values.

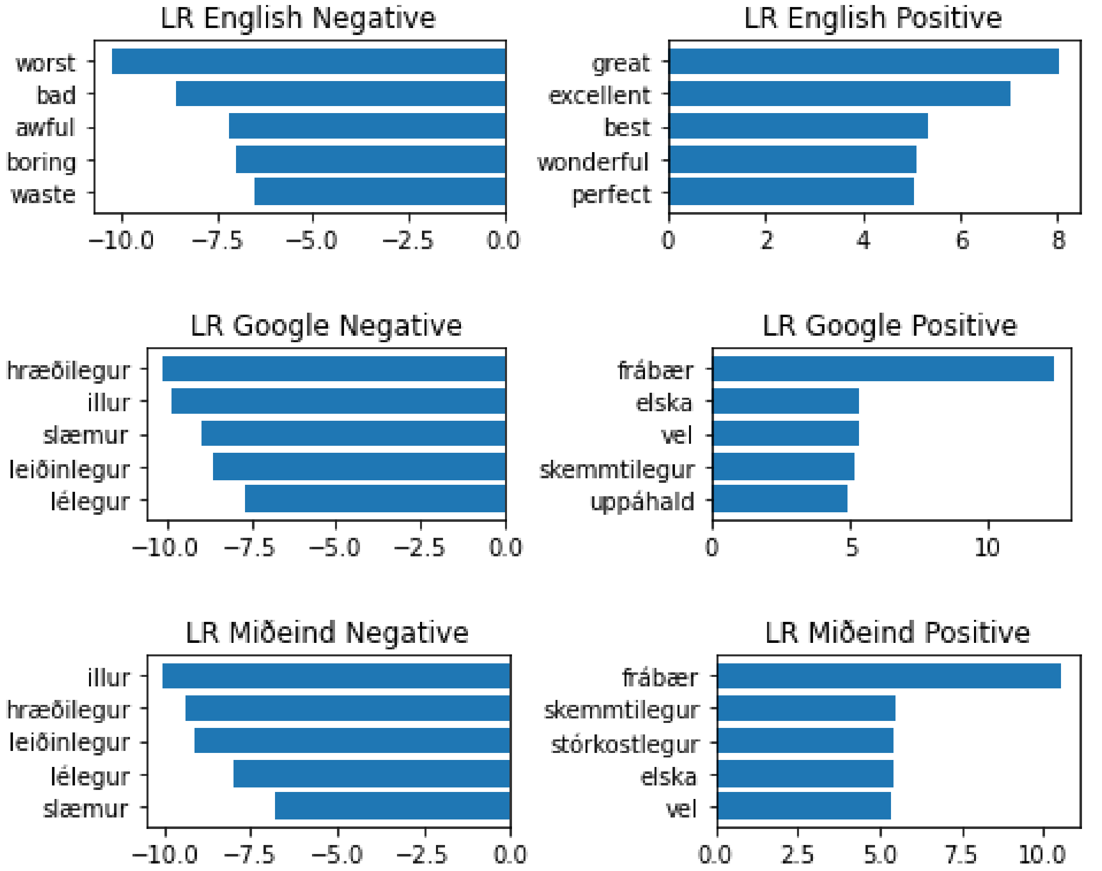
Most important features SVC



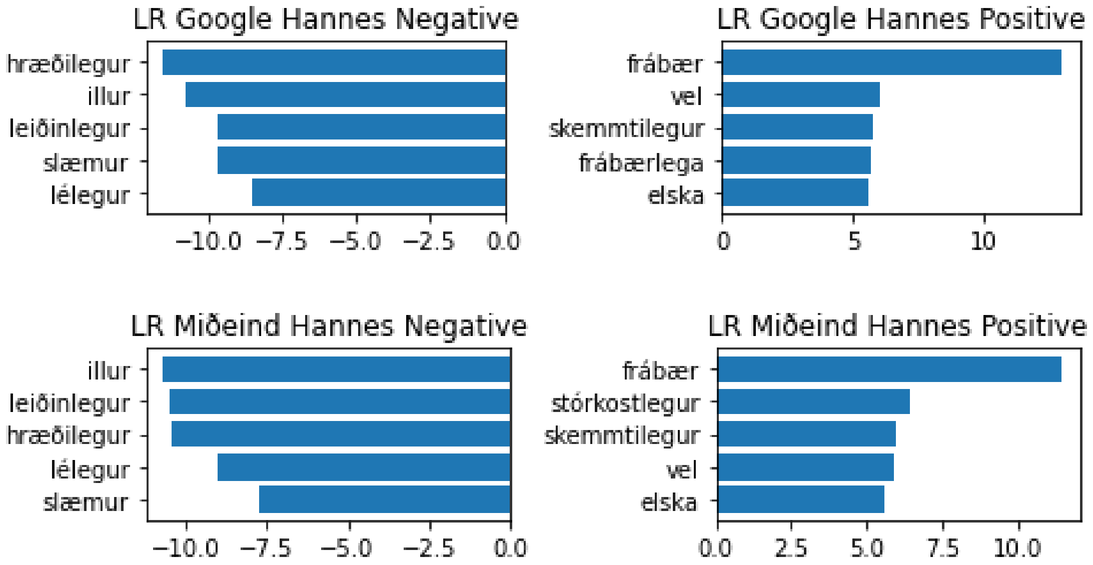
Most important features SVC



Most important features Logistic Regression



Most important features Logistic Regression



2.4 Conclusion of Machine Learning Classifiers

These figures suggest that sentiment analysis can carry across Machine Translation when utilizing state-of-the-art machine learning APIs such as Support Vector Classifier or Logistic Regression. The loss in accuracy during translation is minimal, with 0.6%~ drop in accuracy for Google and 1.5%~ for Miðeind with the IMDB reviews, favoring Google's performance using Support Vector Classifiers.

When using the classifiers on the hand-written reviews from Hannes we noticed that Logistic Regression trained on the dataset from Miðeind Vélþýðing gave the best performance, 45.36% for negative and 89.22% for positive.

This gives us the conclusion that even though Classifiers trained using Google Translated text is best at evaluating it's own test dataset, it seems that Miðeind Vélþýðing trained using Logistic Regression performs the best on text that is hand-written by a native Icelander.

3 Neural Network Models

We utilized transformer based models trained using the RoBERTa-base architecture, for Icelandic we used the IceBERT and Electra models from Huggingface and for English we used RoBERTa base

3.1 Pre-Processing and tokenization

We used the unproccessed dataset because we found that removing stop words that the corpus was not unique enough so that the deep learning model instead of generalizing

we used 4 epochs and learning rate of 1e-6 getLinearSchedule with warmup and adam, we found after 4 epochs that the loss becamse to much indicating that the model was overfitting and therefore memorizing and not learning the data

3.2 Training

3.2.1 IMDB Movie Review Data

3.2.2 Hand written Icelandic Data

3.3 Neural Network Prediction Results

roberta-batch-8-unprocessed-model on IMDB-dataset.csv

RoBERTa English	Precision	Recall	F1-Score
negative	95.75	93.90	94.81
positive	94.99	95.89	94.99

Performance of Icebert and Electra models

Model-[*Train Dataset*] (Sentiment)	Precision	Recall	F1-Score
IceBERT-[*Miðeind*] (negative)	90.76	90.56	90.66
IceBERT-[*Miðeind*] (positive)	90.72	90.92	90.82
IceBERT-[*Google*] (negative)	92.31	91.34	91.83
IceBERT-[*Google*] (positive)	92.18	91.19	91.68
Electra-[*Miðeind*] (negative)	92.44	93.50	**92.97**
Electra-[*Miðeind*] (positive)	93.42	92.36	**92.89**
Electra-[*Google*] (negative)	91.89	93.28	92.58
Electra-[*Google*] (positive)	93.18	91.77	92.47

Transformer-models predicted against the Hannes dataset

Model-[*Train Dataset*] (Sentiment)	Precision	Recall	F1-Score
IceBERT-[*Miðeind*] (negative)	69.23	40.22	50.88
IceBERT-[*Miðeind*] (positive)	89.37	96.56	92.89
IceBERT-[*Google*] (negative)	67.41	33.51	44.77
IceBERT-[*Google*] (positive)	88.35	96.88	92.42
ELECTRA-[*Miðeind*] (negative)	67.44	48.60	**56.49**
ELECTRA-[*Miðeind*] (positive)	90.63	95.49	**92.99**
ELECTRA-[*Google*] (negative)	62.85	49.16	55.17
ELECTRA-[*Google*] (positive)	90.62	94.42	92.48

3.3.1 References

[1] R. Manurung, Franky "Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of EnglishMovie Reviews"

[2] G. Backfried "The Impact of Machine Translation on Sentiment Analysis"

[3] H. Ghorbel, D. Jacot "Sentiment Analysis of French Movie Reviews"

[4] F. Akba, A. Uçan, E. Sezer and H. Sever “ASSESSMENT OF FEATURE SELECTION METRICS FOR SENTIMENT ANALYSES: TURKISH MOVIE REVIEWS”

[5] B. Pang, L. Lee and S. Vaithyanathan “Thumbs up? Sentiment Classification Using Machine Learning Techniques”

[6] Hannes, http://officialstation.com