BSc Final Project
Department of Computer Science

Evaluating Icelandic Sentiment Analysis Models Trained on Translated Data

Ólafur Aron Jóhannsson
olafuraj21@ru.is

Birkir Arndal
birkirh20@ru.is

Eysteinn Örn
eysteinnj19@ru.is

Supervised by Stefán Ólafsson and Hrafn Loftsson

Examined by Sigurjón Ingi Garðarsson

15 December, 2023

Contents

Abstract

In this paper, we experiment with sentiment classification models for Icelandic that leverage machine-translated data for training. We machine translated 50,000 English IMDb reviews, that have been labeled positive and negative, into Icelandic. We evaluate if sentiment effectively carries across after machine translation and, moreover, the accuracy of the classification on native Icelandic text. We analyse the difference between three types of baseline classifiers, Support Vector Machines, Logistic Regression and Naive Bayes, when trained on translated data generated by Google Translate and Miðeind Vélþýðing (e. Translate). Furthermore, we fine-tune and evaluate three pre-trained transformer-based models, RoBERTa, IceBERT and ELECTRA on both the original English texts and the translated texts. Our results indicate that the transformer models perform better than the baseline classifiers on all datasets. Moreover, our evaluation shows that the transformer models can be used to effectively classify sentiment on native Icelandic movie reviews.

1 Introduction

Sentiment analysis uses Natural Language Processing (NLP) to identify, extract, and quantify subjective information in texts, such as positive, negative, or neutral sentiments. It holds utility across various applications, including assessing public opinions, allowing businesses to gauge and categorize customer satisfaction, and offering valuable insights into user-generated content on diverse digital platforms—such as customer reviews, complaints, and comments.

In light of the fact that there are no openly available sentiment analysis models for the Icelandic language, our research takes on the challenge of developing one. We utilize machine learning (ML) techniques for sentimental analysis, a decision motivated by our interest in ML. However, the development of such models typically requires a large corpus of labeled data, which is not readily available for the Icelandic language. To address this obstacle, we have adopted the approach of machine translating into Icelandic an existing labeled English IMDb reviews dataset [1], previously utilized for training sentiment analysis models in English. For other languages, researchers have previously resorted to machine translation (MT) for data generation to address the scarcity of data [2, 3].

This strategy to create an Icelandic sentiment analysis model involves two key methodologies:

1. **Machine-translated datasets**: To address the scarcity of Icelandic sentiment data, we have translated the IMDb dataset into Icelandic using Google Translate and Miðeind Vélþýðing[1] [4, 5]. This approach not only compensates for the lack of native data but also allows us to explore the efficacy of MT in capturing sentiment nuances in Icelandic.
2. **ML model development**: We develop several different ML-based sentiment analysis models, specifically for the Icelandic language, and compare them for best results.

An innovative aspect of our methodology is comparing the the outputs of two MT services: Google Translate and Miðeind Vélþýðing. This will provide insights into which translation tool is more effective for Icelandic sentiment analysis. The model's performance will be validated on actual Icelandic movie reviews to assess its practical utility.

Our research has three primary objectives:

[1] Vélþýðing: https://velthyding.is/

1. **Assessing Sentiment Translation Accuracy**: We aim to investigate if sentiment in English movie reviews is accurately preserved when translated into Icelandic.
2. **Developing Icelandic Sentiment Analysis Resources**: We aim to provide three key resources:
    - An openly available sentiment analysis model for Icelandic movie reviews, addressing the current lack of such tools for the language.
    - Two variations of a machine-translated dataset of 50,000 movie reviews, to serve as a foundational corpus for both our models and future research in Icelandic NLP.
    - To provide an open source pipeline for creating Icelandic machine-translated datasets and models for other domains and tasks.

The IMDb movie review corpus, originally used for training sentiment analysis models in English, provides a comprehensive and diverse source of textual sentiment expressions. By translating this corpus, using both Google Translate and Miðeind Vélþýðing, we aim to compare the effectiveness of a global translation tool against a specialized, localized one in the context of sentiment analysis.

Our hypotheses are as follows:

1. Assuming that meaning is not lost in translation, sentiment classification on native Icelandic text will perform similarly to English.
2. Given that MTs are not perfect, a model trained on the original English dataset will obtain the highest accuracy.
3. Assuming that Miðeind Vélþýðing was created using fewer resources than Google Translate, all of our classifiers trained on MTs from Google Translate will achieve the highest accuracy.
4. Given that IceBERT [5] is pre-trained on the largest Icelandic datasets and assuming that Vélþýðing has more translation errors compared to the more mature Google Translate, sentiment classification on Icelandic text will produce the most optimal results when trained on IceBERT in conjunction with Google Translate.

2 Related Work

Several researchers have used machine learning and deep learning approaches with machine-translated data for sentiment classification. This includes studies using the IMDb dataset, similar to our research, but extending to languages such as Persian, Serbian, Spanish, Russian, French, and Japanese.

Maas *et al.* [1] used a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content. They also introduced a large dataset of movie reviews, the IMDb dataset, to serve as a more robust benchmark for work in sentiment classification[2].

Shalunts *et al.* [6] explored the impact of MT on sentiment analysis, using state-of-the-art tools, SentiSAIL and SDL Language Weaver. The study involved translating original corpora from German, Russian, and Spanish languages, which comprised general news content, into English. They found that the worst case performance decrease was a negligible 5%.

Poncelas *et al.* [7] proposed an MT system in terms of sentiment preservation. They used a dataset consisting of customer feedback in English, French, Spanish, and Japanese. They translated the non-English feedback into English and then classified all the feedback as either positive or negative. They found that the classifiers do not classify translated data as well as original sentences, but that the translation quality is not completely correlated to the performance of the classifier.

Franky and Manurung [3] employed three well known classification techniques, Naive Bayes, Maximum Entropy and Support Vector Machines, with unigram presence and frequency values as the features. Analysis of the Indonesian translations yielded an accuracy of up to 78.82% compared to %80.09 for the original English data.

Dashtipour *et al.* [8] presented a novel, context-aware, deep-learning-driven, Persian sentiment analysis approach and show that LSTM obtained better performance as compared to multilayer perceptron (MLP), autoencoder, Support Vector Machine (SVM), Logistic Regression and CNN algorithms.

Lohar *et al.* [2] present the outcomes of an experiment addressing the complexities inherent in constructing an MT system for user-generated content, specifically tackling the challenges posed by a complex South Slavic language. The focus is directed towards translating English IMDb user movie reviews into Serbian within a low-resource context. The investigation delves into the potentials and limitations of two approaches: (i) phrase-based and (ii) neural MT systems. These systems were trained using out-of-domain clean parallel data sourced from news articles. The primary observations revealed that, even in this low-resource scenario with domain mismatch, the neural approach outperformed the phrase-based approach in handling morphology and syntax.

Amulya *et al.* [9] assessed the accuracy of both ML and Deep Learning (DL) models, trained on the IMDb movie reviews. While ML algorithms operate within a single layer, DL algorithms function across multiple layers, yielding superior outcomes. This study facilitated researchers in discerning the optimal algorithm for sentiment analysis. Comparative analysis between ML and DL approaches showed that DL algorithms exhibit precision and efficiency in results.

Kapukaranov and Nakov [10] presented a system for fine-grained sentiment analysis in Bulgarian movie reviews. They created freely available resources: a dataset of movie reviews with fine-grained

[2] http://ai.stanford.edu/~amaas/data/sentiment/

scores, and a sentiment polarity lexicon. They further compared experimentally the performance of classification, regression and ordinal regression in a 3-way, 5-way and 11-way classification setups, using as features not only the text from the reviews, but also contextual information in the form of metadata, e.g., movie length, director, actors, genre, country, and various scores: IMDB, Cinexio, and user-average. Their results showed that adding contextual information yields strong performance gains.

Qaisar [11] experimented with using Long Short-Term Memory (LSTM) classifier for analyzing sentiments of the IMDb movie reviews. LSTM is a variant of a Recurrent Neural Network (RNN). The data was effectively preprocessed and partitioned to enhance the post classification performance. The results showed a best classification accuracy of 89.9%. The author argued that the results confirm the potential of integrating the designed solution in modern text based sentiments analyzers.

Lee and Kim [12] applied sentiment analysis to the BNC64 corpus data of men's and women's speech. Three different types of analyses were employed: dictionary-based analysis, GRU-based analysis, and BERT-based analysis. When the data were analyzed with the dictionary-based analysis, there was no significant difference in the use of sentiment words between men and women. When the data were analyzed with the GRU-based and BERT-based analysis, it was observed that even though men and women used a similar proportion of sentiment words, women used more positive words. The tendency became much clearer in the BERT-based analysis.

Ghosh [13] employeed three distinct supervised learning methods for sentiment analysis on IMDb reviews: Linear Support Vector Machine, Logistic Regression, and Multinomial Naive Bayes Classifier, each with varied hyperparameter settings. Additionally, the utilization of N-grams was adopted to capture informal jargon nuances. A comprehensive comparative analysis was conducted to determine the optimal model for each supervised learning technique, considering Accuracy Score, F1-Score, and AUC Score. The outcomes of this approach yielded a top accuracy score of approximately 0.910, and a mean F1-score of approximately 0.894 following a 10-fold Cross-validation process.

Though many of these approaches have been successful, these methods are largely under-researched for the Icelandic language. This presents an opportunity to advance NLP for the Icelandic language, particularly in examining how sentiment analysis, when applied through machine-translated content, retains its accuracy and relevance.

3 Methods

Our methodology involved developing sentiment classification models that leverage machine-translated data for training, aiming to reliably predict sentiment in native Icelandic movie reviews. We utilized the IMDb movie review dataset for both training and evaluation. For baseline classifiers, we used Naive Bayes, Support Vector Machine and Logistic Regression provided by Scikit-learn[3], and for advanced models, we utilized pre-trained transformer models such as RoBERTa [14], IceBERT [5], and ELECTRA [15] available from Hugging Face[4].

3.1 Data

Icelandic lacks a dataset for training models for binary sentiment classification, a task that has been extensively carried out in English. We addressed this by translating the English IMDb into Icelandic. The dataset consists of 50,000 reviews, evenly divided into 25,000 positive and 25,000 negative sentiments, categorized by their rating. Reviews with a rating of 4 or below are negative, and those with ratings of 7 and above are positive. The rest were considered neutral and excluded from the dataset.

Table 1 shows two examples of movies reviews from IMDb and their respective sentiment level.

Table 1: English movie reviews with sentiment

Movie Review Text	Sentiment
If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it.Great Camp!!!	Positive
This film contains far too much meaningless violence. Too much shooting and blood. The acting seems very unrealistic and is generally poor. The only reason to see this film is if you like very old cars.	Negative

We also evaluated our sentiment analysis models on real-world data, distinct from our machine-translated dataset. This step provides insight into the effectiveness and applicability of our models trained on translated data in practical scenarios using reviews originally written in Icelandic. For the real-world data, we curated Icelandic movie reviews from two sources:

- 209 reviews from Twitter @kvikmyndaryni account[5].
- 1,111 reviews from officialstation.com, a blog by Hannes Agnarsson Johnson[6].

These reviews had star ratings on a scale from 1 to 10. To align these ratings with the IMDb dataset, we categorized scores of 1–4 as negative and 7–10 as positive. This resulted in a total of 63 negative reviews and 745 positive reviews. To address this imbalance, we selected all 63 negative reviews

[3] https://scikit-learn.org/stable/
[4] https://huggingface.co/
[5] https://twitter.com/kvikmyndaryni
[6] http://officialstation.com

from both datasets and randomly sampled 63 positive reviews to maintain a balance equivalent to that of the IMDb dataset.

When evaluating, we selected the transformer model that had the highest accuracy on machine-translated Icelandic to apply on real-world data, we conducted 10 runs, with each run consisting of a random sample of 50 positive and 50 negative reviews. We calculated the average accuracy of the 10 runs.

We evaluate the accuracy of our models and their ability to adapt to natural language usage using these datasets. This evaluation allows us to explore some of the practical applications of the models.

Table 2 shows two examples of Icelandic movie reviews.

Table 2: Icelandic movie review with sentiment

Movie Review Text	Sentiment
Mögnuð mynd. Intense hljóð og tónlist skapaði mjög dramatíska stemningu. étt keyrsla mikið í gangi og verið að hoppa fram og til baka í mismunandi tímabil. hugaverð saga og persónur. Fullt af geggjuðum leikurum. Virkilega flott mynd enda ekki við öðru að búast frá Christopher Nolan.	Positive
Önnur klisjukennd og fyrirsjáanleg mynd. Ekki gott handrit mikið af vandræðalegum og þvinguðum væmnum atriðum. netflix	Negative

For the baseline classifiers, the data was divided into training and test sets, with 67% (33,500 reviews) allocated for training and 33% (16,500 reviews) reserved for testing the models' accuracy. Conversely, for the transformer models, the data division was adapted to meet their specific needs, which include further splitting the test data into validation and test sets. The validation set is utilized during the model's training phase to fine-tune its parameters, while the test set is used to evaluate the model's accuracy. Accordingly, the dataset was divided into 70% (35,000 reviews) for training, 15% (7,500 reviews) for validation, and 15% (7,500 reviews) for testing.

3.1.1 Translations

We utilized Google Translate and Miðeind Vélþýðing[7] for the MT of the IMDb movie reviews to investigate which MT system more effectively preserves the sentiment. This can be seen by evaluating Icelandic sentiment models trained on the translated data.

The rationale for selecting these tools is twofold. First, Google Translate is known for its wide usage and effectiveness for multiple languages, and it offers a baseline for quality and reliability in translation. Second, in contrast, Miðeind Vélþýðing is a product of Miðeind – a company specializing in NLP and Artificial Intelligence technologies for the Icelandic language – which offers a more localized approach. It uses deep neural networks specifically trained for translating to and from Icelandic, potentially capturing nuances of the language more accurately.

[7] https://huggingface.co/mideind/nmt-doc-en-is-2022-10

Google Translate Utilizes a hybrid model that combines a transformer[16] encoder with an RNN [8] (Recurrent Neural Network) decoder. All the reviews were translated using the `googletrans` Python library, which uses the Google Translate API. The only preprocessing step applied to the raw data was the removal of `<br/>` tags. The absence of errors during the translation process could likely be attributed to the API's maturity and extensive user adoption. Nonetheless, it is noteworthy that the translated reviews occasionally exhibited peculiarities in wording.

Table 3 shows two examples of reviews translated by Google Translate.

Table 3: Translated text using Google Translate (the English text can be seen in Table 1).

Movie Review Text	Sentiment
Ef þér líkar við frumlegan hlátur, muntu líka við þessa mynd. Ef þú ert ungur eða gamall þá muntu elska þessa mynd, helvíti jafnvel mömmu líkaði hana. Frábær búðir!!!	Positive
Þessi mynd inniheldur allt of mikið tilgangslaust ofbeldi. Of mikið skot og blóð. Leikurinn virðist mjög óraunhæfur og er almennt lélegur. Eina ástæðan til að sjá þessa mynd er ef þú hefur gaman af mjög gömlum bílum.	Negative

Miðeind Vélþýðing The Miðeind Vélþýðing translation model uses the multilingual BART [17] model and was trained using the Fairseq sequence modeling toolkit within the PyTorch framework [5]. The Translator encountered challenges when translating the English reviews 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 translated in batches by the Miðeind Vélþýðing.

Additionally, for the resulting Miðeind machine-translated dataset, it was necessary to remove lengthy nonsensical words (e.g., "...BARNABARNABARNAÞÁTTURINN"), and reduce sequences of the same character repeated in succession into a single character (e.g., "jáááááááá" to "já").

Table 4 shows two examples of reviews translated by Miðeind Vélþýðing.

[8] https://www.ibm.com/topics/recurrent-neural-networks

Table 4: Translated text using Miðeind Vélþýðing (the original text can be seen in Table 1).

Movie Review Text	Sentiment
Ef þú ert hrifin/n af skrækjandi hlátri úr maganum á þér mun þér líða vel í þessari mynd. Hvort sem þú ert ung eða gömul muntu verða hrifin/n af þessari mynd, jafnvel mamma hafði gaman af henni. Frábærar búðir!	Positive
Í þessari mynd er allt of mikið af tilgangslausu ofbeldi. Of mikið af skotbardögum og blóði. Leikurinn virðist óraunhæfur og yfirleitt frekar lélegur. Eina ástæðan fyrir því að þú vilt sjá þessa mynd er sú að þú ert hrifinn af mjög gömlum bílum.	Negative

3.2 Baseline classifiers

Our baseline classifiers are a set of established algorithms that serve as a starting point for model performance evaluation. The accuracy of these classifiers establishes a minimum threshold that the more complex models should exceed.

We selected the following classifiers as our baseline:

- **Logistic Regression**: This statistical algorithm is used to predict the probability that a given input belongs to a certain class. It employs a logistic function to estimate the likelihood of a class, which in our context is categorized as either positive or negative.
- **Multinomial Naive Bayes Classifier**: Naive Bayes (NB) is collection of algorithms based on Bayes' theorem that assumes all features are mutually independent within a given a class. Multinomial Naive Bayes is a variant of NB which assumes that the feature probabilities follow a multinomial distribution.
- **Linear Support Vector Classification**: A variant of Support Vector Machine (SVM) that aims to find the optimal separating hyperplane, thereby maximizing the margin between two distinct classes.

The input to the classifiers was data in the form of term frequencies, calculated using the TF-IDF (term frequency-inverse document frequency[9]) vectorizer from Scikit-learn.[10] This allows the classifiers to weigh the importance of a each term in the corpus relative to its frequency across the entire dataset.

3.2.1 Normalization

After translation, we took various steps to transform and process the text for the classifiers, such as removing erroneous and irrelevant data. Compared to the transformer models, the baseline classifiers required more extensive data transformations.

Before beginning data normalization, the first step was tokenization. For the original English

[9] https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

[10] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

dataset, we used a tokenizer from the Natural Language Toolkit[11] (NLTK). In contrast, for the machine-translated datasets, we utilized a tokenizer[12] specifically designed for Icelandic, provided by Miðeind.

The normalization steps for the baseline classifiers were as follows:

- **Remove Noise**: Brackets, HTML tags, and certain special characters were removed. Punctuation was also removed, except in the case of abbreviations, to reduce noise in the data.
- **Sentiment Conversion**: The sentiment labels were changed to a binary format, with 0 for negative and 1 for positive.
- **Lowercasing**: This step normalized and reduced the vocbulary of the datasets by converting all texts to lowercase.
- **Remove Stop Words**[13]: Filler words that do not contribute significantly to the meaning of the sentences were removed, which improved the accuracy of the classifiers.[14] The stop words used came from Árni Magnússon Institute for Icelandic Studies. [15]
- **Lemmatization**: Different forms of the same word were converted to a standardized form, reducing the datasets' vocabulary and improving the classifiers' accuracy.[16]
- **Mark Negation**: Text following a negation word and up to a punctuation mark was suffixed with `_NEG`. This helped the classifiers understand sentence context by marking the scope of negation. Our analysis indicated that this approach improved the accuracy of the classifiers.

We developed a custom normalization class in Python to execute all the normalization steps above, with the exception of lemmatization. For lemmatization, we employed a rule-based lemmatizer for Icelandic text known as Nefnir[17] [18]. Nefnir needs part-of-speech tagged text, for which we used IceStagger [19], which is part of the IceNLP toolkit[18].

Table 5 and 6 show two examples of normalized reviews by Google Translate and Miðeind Vélþýðing.

Table 5: Normalized Icelandic movie review with sentiment translated using Google Translate.

Movie Review Text	Sentiment
líka frumlegur hlátur muna líkur mynd vera ungur gamall muna elska mynd helvíti jafnvel mamma líka hana.frábær búð	Positive
vera leiðinlegur atriði þrúgandi dimmur mynd reyna lýsa konar siðferði falla boðskapur sinn endurleysandur eiginleiki ofanálag halda geta ekki láta_NEG bókaverð_NEG líta_NEG mikill_NEG vera_NEG óglamorískur_NEG gera_NEG	Negative

[11] https://www.nltk.org/
[12] https://github.com/mideind/Tokenizer
[13] https://github.com/atlijas/icelandic-stop-words
[14] https://kavita-ganesan.com/what-are-stop-words/
[15] https://arnastofnun.is/en/about
[16] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
[17] https://github.com/jonfd/nefnir
[18] https://github.com/hrafnl/icenlp

Table 6: Normalized Icelandic movie review with sentiment translated using Miðeind Vélþýðing

Movie Review Text	Sentiment
vera hrífa skrækjandi hlátur magi munu líða vel mynd vera ungur gamall muna verða hrífa mynd jafnvel mamma hafa gaman hún frábær búð	Positive
vera leiðinlegur atriði kúga myrkur kvikmynd reyna draga konar siðferðislegur mynd falla flatt boðskapur sinn eiginleiki safna upp ofanálag halda geta ekki gera_NEG bókaverð_NEG ógeðfelldur_NEG það_NEG	Negative

3.3 Transformer models

A transformer model is a type of artificial neural network recognized for it's multi-head attention mechanism and absence of recurrent units. It stands out for it's faster training compared to prior models like LSTM (Long Short-Term Memory [19]). The transformer model employs a mechanism called self-attention to understand the context within a sequence of data [16]. The transformer model architecture gave rise to pre-trained systems like GPTs and BERT. The specific transformer models that we utilized are:

- **RoBERTa**: An enhanced version of BERT [20], pre-trained on 160 GB of English textual data. We employed RoBERTa base[20] model for training on the original English IMDb dataset.
- **IceBERT**[21] [5]: A variant of the RoBERTa model developed by Miðeind, pre-trained on a combination of the Icelandic Gigaword Corpus (IGC) [21] and web data, 15.8 GB in total.
- **ELECTRA**: A transformer model that implements an innovative pre-training approach by simultaneously training two distinct transformer models: a generator and a discriminator. The generator changes tokens to fake tokens, while the discriminator predicts which tokens have been changed by the generator. We used the electra-base-igc-is[22] model [22], which was pre-trained on the IGC, encompassing 8.2 GB of Icelandic textual data.

RoBERTa and IceBERT tokenize using the Byte Pair Encoding method (BPE)[23], while ELECTRA uses the WordPiece[24] method.

3.3.1 Normalization

Sentiment labels were changed to a binary format for all datasets. For the translated datasets, noise removal was necessary prior to tokenization, similar to the 'Remove Noise' step performed for the baseline classifiers. This step is crucial because translation may introduce errors or irrelevant information not present in the original dataset, which could potentially impair the model's accuracy.

Conversely, the English dataset required no further normalization before tokenization. Our observations indicated that transformer models yield better results when trained on more diverse corpora,

[19] https://www.researchgate.net/publication/13853244_Long_Short-term_Memory
[20] https://huggingface.co/roberta-base
[21] https://huggingface.co/mideind/IceBERT
[22] https://huggingface.co/jonfd/electra-base-igc-is
[23] https://huggingface.co/docs/transformers/main/tokenizer_summary#byte-level-bpe
[24] https://huggingface.co/docs/transformers/main/tokenizer_summary#wordpiece

thereby eliminating the need for lemmatization, negation marking, and stop word removal.

3.4 Model training

For our baseline classifiers, we kept the default parameters from the scikit-learn library[25]. The classes and their default parameters can be seen in Table 7.

Classifier	Parameters
MultinomialNB	alpha=1.0, fit_prior=True, class_prior=None
LinearSVC	penalty='l2', loss='squared_hinge', dual=True, tol=0.0001, C=1.0, multi_class='ovr', fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000
Logistic Regression	penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None

Table 7: Default Parameters for Baseline Classifiers

For our transformer models' training, we utilized the AdamW optimizer [23], which is a variant of the standard Adam optimizer [24]. It alters the weight decay application process, effectively decoupling it from the gradient update, which enhances regularization and helps prevent overfitting. We started with an initial learning rate of 1e-6 and used a linear decay schedule, gradually reducing the learning rate to zero throughout the training period. The models underwent training for 4 epochs with a batch size of 8. We observed that extending training beyond this point led to overfitting, as evidenced by an increase in validation loss while the training loss decreased. All transformer model training was executed on an ASUS ROG Strix GeForce RTX™ 3080 graphics card, using CUDA 11.8, Python 3.10 and PyTorch 2.0.1.

3.5 Evaluation method

When assessing the performance of the classifiers, we used the accuracy as our benchmark metric[26]. Accuracy is calculated as follows:

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

Where:

- True Positive (TP) refers to correctly identified positive sentiments.
- False Positive (FP) signifies incorrectly identified positive sentiments.
- True Negative (TN) denotes correctly identified negative sentiments.

[25] https://scikit-learn.org
[26] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

- False Negative (FN) represents incorrectly identified negative sentiments.

4 Results and Analysis

The ELECTRA model achieved an accuracy of 92.24% trained on Google-translated data and 92.16% for Miðeind-translated data. The accuracy on the Icelandic natural data, where we randomly sampled fifty positive and fifty negative reviews 10 times based on seed, was 90.9% and 91.5% for Miðeind Vélþýðing and Google Translate, respectively, with the latter slightly outperforming the former. For comparison, the RoBERTa model, trained and evaluated on the original English data, obtained an accuracy of 94.9%.

4.1 Baseline Classifiers

The accuracy of each baseline classifier trained on the English dataset and the machine-translated datasets are shown in Table 8 and Table 9, respectively. The best-performing baseline classifier was the Support Vector Classifier, which achieved an accuracy of 89.02% on the Google-translated dataset. This compares to an accuracy of 89.68% for the SVC trained on the original English data, representing a small difference in accuracy of 0.66%.

Classifier	**Accuracy**
SVC	**89.68%**
NB	85.79%
LR	89.35%

Table 8: Classifier accuracy on the original English IMDb dataset

Classifier	**Accuracy**	
	Google	**Miðeind**
SVC	**89.02**%	88.15%
NB	85.78%	85.16%
LR	88.74%	87.76%

Table 9: Classifier accuracy on the translated datasets

4.2 Transformer Models

The transformer model RoBERTa demonstrated the highest accuracy on the original IMDb dataset, with a score of 94.9%, as detailed in Table 10. When we turn our attention to the translated datasets, as shown in Table 11, ELECTRA emerges as the best performer with an accuracy of 92.24% on the Google-translated data. Compared to the RoBERTa model, which obtained an accuracy of 94.9% when trained and evaluated on the original English data, the drop in performance of translated data is 2.66%.

Classifier	**Accuracy**
RoBERTa	94.9%

Table 10: Transformer accuracy on the original English IMDb dataset

Classifier	Accuracy	
	Google	**Miðeind**
IceBERT	92.18%	90.74%
Electra	**92.24**%	92.16%

Table 11: Transformer accuracy on the translated datasets

4.3 Icelandic Reviews

We evaluated the best performing model, trained on translated data (i.e. ELECTRA), on the movie reviews originally written in Icelandic. We ran the evaluation 10 times with 100 sampled reviews split evenly with 50 positive and 50 negative reviews, and averaged the accuracy. The results, detailed in Table 12 and 13, show that ELECTRA trained on Google-translated data and Miðeind-translated data obtained an accuracy of 91.5% and 90.9%, respectively.

Table 12: Icelandic Dataset translated by Miðeind Vélþýðing, evaluation using ELECTRA

Run	Accuracy	Seed
1	94%	1570
2	91%	3822
3	89%	3964
4	89%	950
5	91%	16
6	94%	2287
7	91%	4026
8	90%	3741
9	90%	2897
10	90%	4232
Average	90.9%	
Std. Deviation	1.69%	

Table 13: Icelandic Dataset translated by Google, evaluation using ELECTRA Google

Run	Accuracy	Seed
1	91%	7936
2	93%	5661
3	90%	1143
4	93%	4007
5	90%	688
6	93%	4523
7	92%	2195
8	92%	9617
9	89%	2329
10	92%	6348
Average	91.5%	
Std. Deviation	1.36%	

5 Discussion

Our work outlines a methodology for creating data and machine learning models for sentiment analysis of Icelandic movie reviews. Our findings indicate that this task is feasible using current state-of-the-art ML methods and NLP tools.

Our first hypothesis was that sentiment classification on native Icelandic text would perform similarly to English. Our findings suggest that employing sentiment classification models trained on machine-translated Icelandic yields performance very similar to models trained on the original English data – the drop in accuracy is only a 2.66%.

We found support for our second hypothesis, which was that models trained on the original English dataset will obtain the highest accuracy, given that the original texts used for the model training is in English. We found that the RoBERTa model trained on English data did perform the best of all the models, obtaining an accuracy of 94.9%.

The most accurate baseline model was the Support Vector Classifier, trained using Google-translated data, with an accuracy of 89.02%. The most accurate neural network model was ELECTRA, trained using Google-translated data, with an accuracy of 92.24%. Comparatively, the RoBERTa model, which is trained on original English data, achieved an accuracy of 94.9% – thus, the drop in accuracy is 2.66%. Our third hypothesis is that models trained on Google-translated data would obtain the highest accuracy was correct across all of the models. We also hypothesized that IceBERT (a RoBERTa model) would be the best transformer-model. We did not find support for this claim, since the ELECTRA model obtained the highest accuracy on the translated data.

We also note that the accuracy is similar when evaluating the model on Icelandic natural language data. ELECTRA, trained using Miðeind-translated data, achieved an average of 90.9% and trained using Google-translated data the same model obtained an average accuracy of 91.5%,

We observed that syntactically the text from both Miðeind and Google are most often translated correctly, and that the semantic meaning of the text in both cases transfers when we do sentiment evaluation on the translations, even though the syntax is different between translators. It could be argued that although Google's model is more mature, Miðeind's translations tend to yield more natural-sounding Icelandic text. Conversely, Google's translation are often more literal and closer to the verbatim content of the original data, potentially resulting in translations that may feel less natural.

When developing a sentiment classification model, the ease of adoption of Support Vector Classifiers, combined with their excellent performance, is a metric that should be considered. ELECTRA performs better then the baseline, and could potentially achieve even better results than our indicated findings if trained with a larger corpus, more epochs, or different hyperparameters. It could possibly reach a performance level similar to our RoBERTa model which was trained on English IMDb data, around 95%.

These findings are promising for the task of sentiment analysis in Icelandic for other domains where such information is useful, such as for customer reviews or company valuations.

5.1 Limitations

Our research has several limitations. The first limitations revolved around time constraints and computing power. Training transformer models can be time-consuming and resource-intensive, but this is contingent on the dataset provided for the model. Finally, our methodology may not

generalize to other domains beyond sentiment classification on movie reviews. Other domains and tasks require bespoke approaches to data collection and processing, as well as modeling methods.

5.2 Future work

The opportunity to conduct additional downstream training on the models allows for fine-tuning them specifically for sentiment classification in a given use case. Alternatively, one could explore the feasibility of employing our methodology, which involves translating data and utilizing it to train models for various classification tasks in Icelandic.

References

[1] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 142–150. [Online]. Available: https://aclanthology.org/P11-1015

[2] P. Lohar, M. Popović, and A. Way, "Building English-to-Serbian Machine Translation System for IMDb Movie Reviews," in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, T. Erjavec, M. Marcińczuk, P. Nakov, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, and R. Yangarber, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 105–113. [Online]. Available: https://aclanthology.org/W19-3715

[3] Franky and R. Manurung, "Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews," 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:18531056%7D

[4] H. B. Símonarson, H. P. Jónsson, P. O. Ragnarsson, S. L. Ingólfsdóttir, V. Þorsteinsson, and V. Snæbjarnarson, "Long Context Translation Models for English-Icelandic translations (22.09)," 2022, CLARIN-IS. [Online]. Available: http://hdl.handle.net/20.500.12537/278

[5] V. Snæbjarnarson, H. B. Símonarson, P. O. Ragnarsson, S. L. Ingólfsdóttir, H. Jónsson, V. Thorsteinsson, and H. Einarsson, "A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 4356–4366. [Online]. Available: https://aclanthology.org/2022.lrec-1.464

[6] G. Shalunts, G. Backfried, and N. Commeignes, "The Impact of Machine Translation on Sentiment Analysis," in *2016 The Fifth International Conference on Data Analytics*. Marseille, France: IARIA, 2016, pp. 4356–4366, available for download: https://www.thinkmind.org/download.php?articleid=data_analytics_2016_3_20_60032.

[7] A. Poncelas, P. Lohar, J. Hadley, and A. Way, "The Impact of Indirect Machine Translation on Sentiment Classification," in *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, M. Denkowski and C. Federmann, Eds. Virtual: Association for Machine Translation in the Americas, Oct. 2020, pp. 78–88. [Online]. Available: https://aclanthology.org/2020.amta-research.7

[8] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment Analysis of Persian Movie Reviews Using Deep Learning," *Entropy*, vol. 23, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235204861

[9] K. Amulya, S. B. Swathi, P. Kamakshi, and D. Y. Bhavani, "Sentiment Analysis on IMDB Movie Reviews using Machine Learning and Deep Learning Algorithms," *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 814–819, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247110030

[10] B. Kapukaranov and P. Nakov, "Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, R. Mitkov, G. Angelova, and K. Bontcheva, Eds. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sep. 2015, pp. 266–274. [Online]. Available: https://aclanthology.org/R15-1036

[11] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–4, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227220638

[12] Y.-H. Lee and J.-H. Kim, "A Sentiment Analysis of Men's and Women's Speech in the BNC64," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, K. Hu, J.-B. Kim, C. Zong, and E. Chersoni, Eds. Shanghai, China: Association for Computational Linguistics, 11 2021, pp. 603–610. [Online]. Available: https://aclanthology.org/2021.paclic-1.63

[13] A. Ghosh, "Sentiment Analysis of IMDb Movie Reviews : A Comparative Study on Performance of Hyperparameter-tuned Classification Algorithms," *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 289–294, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249475557

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, cite arxiv:1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[15] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in *ICLR*, 2020. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[18] S. L. Ingólfsdóttir, H. Loftsson, J. F. Daðason, and K. Bjarnadóttir, "Nefnir: A high accuracy lemmatizer for Icelandic," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, M. Hartmann and B. Plank, Eds. Turku, Finland: Linköping University Electronic Press, Sep.–Oct. 2019, pp. 310–315. [Online]. Available: https://aclanthology.org/W19-6133

[19] H. Loftsson and R. Östling, "Tagging a Morphologically Complex Language Using an Averaged Perceptron Tagger: The Case of Icelandic," in *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, S. Oepen, K. Hagen, and J. B. Johannessen, Eds. Oslo, Norway: Linköping University Electronic Press, Sweden, May 2013, pp. 105–119. [Online]. Available: https://aclanthology.org/W13-5613

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[21] S. Steingrímsson, S. Helgadóttir, E. Rögnvaldsson, S. Barkarson, and J. Guðnason, "Risamálheild: A Very Large Icelandic Text Corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1690

[22] J. F. Daðason and H. Loftsson, "Pre-training and Evaluating Transformer-based Language Models for Icelandic," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7386–7391. [Online]. Available: https://aclanthology.org/2022.lrec-1.804

[23] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference for Learning Representations, San Diego*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980