Evaluating Icelandic Sentiment Analysis Models Trained on Translated Data

Ólafur A. Jóhannsson, Birkir H. Arndal, Eysteinn Ö. Jónsson, Stefán Ólafsson, Hrafn Loftsson
Department of Computer Science
Reykjavik University, Iceland
{olafuraj21, birkirh20, eysteinnj19, stefanola, hrafn}@ru.is

Abstract

We experiment with sentiment classification models for Icelandic that leverage machine-translated data for training. Since no large sentiment dataset exists for Icelandic, we translate 50,000 English IMDb reviews, classified either as positive or negative, into Icelandic using two services: Google Translate and GreynirTranslate. After machine translation, we assess whether the sentiment of the source language text is retained in the target language. Moreover, we evaluate the accuracy of the sentiment classifiers on non-translated Icelandic text. The performance of three types of baseline classifiers is compared, i.e., Support Vector Machines, Logistic Regression and Naive Bayes, when trained on translated data generated by either translation service. 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. Furthermore, our evaluation shows that the transformer models trained on data translated from English reviews can be used to effectively classify sentiment on non-translated Icelandic movie reviews.

Keywords: sentiment classification, movie reviews, machine translation, machine learning

1. Introduction

Sentiment analysis is the task of using Natural Language Processing (NLP) to identify, extract, and quantify subjective information in texts, such as positive, negative, or neutral sentiments. This task has been found to be practically beneficial, both for businesses to understand customer opinions in large volumes of text, e.g., to guide marketing strategies and guide investment decisions (Hartmann et al., 2023), and for research, e.g., analyzing human behavior in social networks (Ramírez-Tinoco et al., 2018), and patient outcomes based on medical records data (Denecke and Deng, 2015).

For the Icelandic language, neither open sentiment analysis models exist nor a large corpus of labelled sentiment data, which is typically required for training such models. For other languages, researchers have previously resorted to machine translation to address data scarcity (Shalunts et al., 2016; Lohar et al., 2019; Poncelas et al., 2020).

Our method to create sentiment analysis models for Icelandic involves two phases:

1. **Machine Translation (MT) of the IMDb dataset**: We use Google Translate and GreynirTranslate[1] (Snæbjarnarson et al., 2021) to machine translate the English IMDb reviews dataset (Maas et al., 2011a) into Icelandic. This approach not only compensates for the lack of Icelandic sentiment data, but also allows us to explore the efficacy of MT in capturing sentiment nuances in Icelandic. By using both Google Translate and GreynirTranslate, we aim to compare the effectiveness of a general-purpose translation tool (from Google) against a specialized, localized one (see Section 3.1.1) in the context of sentiment analysis.

2. **Machine Learning (ML) model development**: We develop and evaluate several different ML-based sentiment analysis models, specifically for the Icelandic language. The set of ML models consist of i) baseline classifiers based on Support Vector Machines, Logistic Regression, and Naive Bayes, as well as ii) the transformer-based models RoBERTa (Liu et al., 2019), IceBERT (Snæbjarnarson et al., 2022), and ELECTRA (Clark et al., 2020) pre-trained on Icelandic data (Daðason and Loftsson, 2022). Furthermore, we validate the model's performance on a small set of movie reviews written in Icelandic.

Our research has two primary objectives:

1. **Assessing Sentiment Translation Accuracy**: We investigate if sentiment in English movie reviews is accurately preserved when translated into Icelandic.

2. **Developing Icelandic Sentiment Analysis Resources**: We provide three key resources:

   - An open sentiment analysis model for Ice-

[1] https://velthyding.is/

21-22 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

landic movie reviews, addressing the current lack of such tools for the language[2].

- 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[3].
- An open source pipeline for creating Icelandic machine-translated datasets and models for other domains and tasks[4].

Our hypotheses are as follows:

1. Assuming that meaning is not lost in translation, sentiment classification on Icelandic text, that have been translated from English, will perform similarly to sentiment classification in English. However, given that MTs are not perfect, models trained on the original English dataset will obtain a somewhat higher accuracy than models for Icelandic, trained on translated data.

2. Provided that that GreynirTranslate was created using fewer resources compared to Google Translate, all of our classifiers trained on data translated by Google Translate will achieve the highest accuracy.

3. Given that IceBERT is pre-trained on the largest Icelandic datasets (Snæbjarnarson et al., 2022) and assuming that GreynirTranslate has more translation errors compared to the more established Google Translate, sentiment classification on Icelandic text is expected to yield the highest accuracy when IceBERT is fine-tuned on translated data generated by Google Translate.

2. Related Work

Maas et al. (2011b) introduced a large dataset of movie reviews, the IMDb dataset (Maas et al., 2011a), to serve as a benchmark for work in sentiment classification. They used a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content. They built a probabilistic model of documents using the word vectors and used a logistic regression classifier for sentiment classification. Their model obtained an accuracy of 87.3–88.9% using a variety of features when evaluated on a test set of 25,000 reviews. The IMDb dataset has provided a standardized benchmark for testing sentiment analysis algorithms and has been influential in advancing research in this area.

Research has shown that it is possible to preserve sentiment post-machine translation from various European languages to English. Shalunts et al. (2016) explored the impact of MT on sentiment analysis, using state-of-the-art tools, SentiSAIL (for sentiment analysis) and SDL Language Weaver (for MT). The study involved translating original corpora from German, Russian, and Spanish, which comprised general news content, into English. They found that the worst case performance decrease in sentiment classification in English was within 5%.

Poncelas et al. (2020) 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 accuracy of the classifier.

Lohar et al. (2019) presented 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 morphologically complex South Slavic language. The focus was directed towards translating English IMDb user movie reviews into Serbian within a low-resource context. The investigation delved 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. (2022) assessed the accuracy of both classical ML models 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.

Researchers have developed sentiment analysis resources for low-resource languages. Kapukaranov and Nakov (2015) presented a system for fine-grained sentiment analysis in Bulgarian movie reviews. They created freely available resources: (i) a dataset of movie reviews with fine-grained scores, (ii) and a sentiment polarity lexicon. They further compared experimentally the performance

[2] https://huggingface.co/Birkir/electra-base-igc-is-sentiment-analysis

[3] https://github.com/cadia-lvl/sentiment-analysis/tree/main/Datasets

[4] https://github.com/cadia-lvl/sentiment-analysis

of classification and regression, using as features the text from the reviews and the 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. Shode et al. (2023) created a dataset of reviews about Nigerian movies. Professional translators translated about 1,000 reviews, originally written in English, to four Nigerian languages, resulting in a multilingual parallel sentiment corpus. The authors train and evaluate both classical machine learning methods and pretrained language models.

Experiments have shown that Deep Neural Networks (DNNs) can effectively model sentiment analysis. Qaisar (2020) experimented with using Long Short-Term Memory (LSTM) classifier for analyzing sentiments of the IMDb movie reviews. 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.

Linear models have also been successfully used for sentiment classification. Ghosh (2022) employed 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 using Linear SVM, and a mean F1-score of approximately 0.894 following a 10-fold cross-validation process.

Though many of these approaches have been successful, they are largely under-researched for the Icelandic language. This presents an opportunity to advance NLP for Icelandic, 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 non-translated 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 as implemented in the Scikit-learn Python library[5]. For advanced models, we utilized the pre-trained transformer models RoBERTa (Liu et al., 2019), IceBERT, which is based on the RoBERTa architecure and pre-trained on Icelandic data (Snæbjarnarson et al., 2022), and a version of ELECTRA (Clark et al., 2020), also pre-trained on Icelandic data (Daðason and Loftsson, 2022) (see Section 3.3).

3.1. Data

Icelandic lacks a dataset for training models for sentiment classification. We addressed this by translating the English IMDb datset 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 remaining reviews were considered neutral and excluded from the dataset. Table 1 shows two examples of movie reviews written in English from IMDb and their respective sentiment level.

We also evaluated our sentiment analysis models on non-translated Icelandic data, distinct from the 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 non-translated data, we curated Icelandic movie reviews from two sources:

- 209 reviews from Twitter @kvikmyndaryni account[6].
- 1,111 reviews from officialstation.com, a blog by Hannes Agnarsson Johnson[7].

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 and to maintain a balance equivalent to that of the IMDb dataset, we selected all 63 negative reviews from both datasets and randomly sampled 63 positive reviews. Table 2 shows two examples of non-translated Icelandic movie reviews.

When evaluating the accuracy on non-translated data, we selected the transformer model that obtained the highest accuracy on machine-translated Icelandic. We conducted 10 runs, with each run consisting of a random sample of 50 positive and 50 negative reviews, which were sampled from the

[5] https://scikit-learn.org/

[6] https://twitter.com/kvikmyndaryni

[7] http://officialstation.com

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
Besides being boring, the scenes were oppressive and dark. The movie tried to portray some kind of moral, but fell flat with its message. What were the redeeming qualities?? On top of that, I don't think it could make librarians look any more unglamorous than it did.	Negative

Table 1: English IMDb movie reviews with sentiment.

Movie Review Text	Sentiment
Mögnuð mynd. Intense hljóð og tónlist skapaði mjög dramatíska stemningu. þétt keyrsla mikið í gangi og verið að hoppa fram og til baka í mismunandi tímabil. áhugaverð 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

Table 2: Non-translated, original Icelandic movie reviews with sentiment.

63 negative and 63 positive reviews, mentioned above.

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. For the transformer models, the test data was further split into validation and test sets. 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 GreynirTranslate (Snæbjarnarson et al., 2021) for the MT of the IMDb movie reviews to investigate which MT system more effectively preserves sentiment. This can be seen by evaluating Icelandic sentiment models trained on data translated by Google Translate, on the one hand, and by GreynirTranslate, on the other.

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, GreynirTranslate is a product of Miðeind[8] – a company specializing in NLP and Artificial Intelligence technologies for the Icelandic language – which offers a more localized approach. It uses DNNs specifically trained for translating to and from Icelandic, potentially capturing nuances of the language more accurately.

Google Translate Utilizes a hybrid model that combines a transformer (Vaswani et al., 2017) encoder with a Recurrent Neural Network (RNN) decoder. All the reviews were translated using the `googletrans` Python library, which uses the Google Translate API[9]. 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.

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

GreynirTranslate Uses the multilingual BART (Lewis et al., 2020) model and was trained using the Fairseq sequence modeling toolkit within the PyTorch framework. The GreynirTranslate model achieved a BLEU score of 24.3 on the English-Iceland news translation task at WMT 2021 (Símonarson et al., 2021). 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 GreynirTranslate.

Additionally, for the resulting machine-translated dataset by GreynirTranslate, it was necessary to remove lengthy nonsensical words (e.g., "...BARNABARNABARNAÞÁTTURINN"), and convert repeated sequences of the same character into a single character (e.g., "jáááááá" to "já").

Table 4 shows two examples of reviews translated by GreynirTranslate.

[8] `https://mideind.is/`

[9] To the best of our knowledge, evaluation results for English-Icelandic translations have not been published.

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
Fyrir utan að vera leiðinleg voru atriðin þrúgandi og dimm. Myndin reyndi að lýsa einhvers konar siðferði, en féll niður með boðskap sínum. Hverjir voru endurleysandi eiginleikarnir?? Í ofanálag held ég að það gæti ekki látið bókaverði líta meira út fyrir að vera óglamorískur en það gerði.	Negative

Table 3: Translated text using Google Translate (the original English 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
Auk þess að vera leiðinleg voru atriðin kúgandi og myrk. Kvikmyndin reyndi að draga upp einhvers konar siðferðislega mynd en féll flatt með boðskap sínum. Hvaða eiginleikar voru það sem söfnuðust upp? Í ofanálag held ég að það gæti ekki gert bókaverði ógeðfelldari en það.	Negative

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

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 is the 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 vectorizer from Scikit-learn. 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

Before beginning text normalization – the process of transforming text into a single canonical form – tokenization is needed. For the original English dataset, we used a tokenizer from the Natural Language Toolkit (NLTK)[10]. In contrast, for the machine-translated datasets, we utilized a tokenizer (Þorsteinsson et al., 2022) specifically designed for Icelandic.

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 vocabulary of the datasets by converting all texts to lowercase.

- **Remove Stop Words**: Stop words (Jasonarson, 2018) that do not contribute significantly to the meaning of the sentences were removed, which improved the accuracy of the classifiers.

- **Lemmatization**: Different forms of the same word were converted to a standardized form, reducing the datasets' vocabulary and improving the classifiers' accuracy.

[10] https://www.nltk.org/

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

Table 5: A normalized version of the movie review from Table 3 that had been translated to Icelandic by Google Translate.

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

Table 6: A normalized version of the movie review from Table 4 that had been translated to Icelandic by GreynirTranslate.

- **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 Nefnir (Daðason, 2017), a rule-based lemmatizer for Icelandic text (Ingólfsdóttir et al., 2019). Nefnir needs part-of-speech tagged text, for which we used IceStagger (Loftsson and Östling, 2013), which is part of the IceNLP toolkit (Loftsson, 2009).

Table 5 and 6 show two examples of normalized reviews translated by Google Translate and GreynirTranslate.

3.3. Transformer Models

A transformer model is a type of neural network characterized by its multi-head attention mechanism and absence of recurrent units. The transformer model employs a mechanism called self-attention for creating contextual embeddings of the input text to understand the context within a sequence of data (Vaswani et al., 2017). The specific transformer models that we utilized are:

- **RoBERTa** (Liu et al., 2019): An enhanced version of BERT (Devlin et al., 2019), pre-trained on 160 GB of English textual data. We fine-tuned the RoBERTa base model (FacebookAI, 2019) on the original English IMDb dataset.
- **IceBERT** (Snæbjarnarson et al., 2022): A variant of the RoBERTa model developed by Miðeind (Miðeind, 2022), pre-trained on a combination of the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018) and web data, 15.8 GB in total.
- **ELECTRA** (Clark et al., 2020): A transformer model that simultaneously trains two distinct transformer models: a generator and a discriminator. The generator turns existing tokens into fake tokens, while the discriminator predicts which tokens have been changed by the generator. We used the Icelandic ELECTRA-base model (Daðason, 2022), which was pre-trained on the IGC, encompassing 8.2 GB of Icelandic textual data (Daðason and Loftsson, 2022).

RoBERTa and IceBERT tokenize the text using the Byte Pair Encoding method (BPE)[11], while ELECTRA uses the WordPiece[12] method.

3.3.1. Normalization

Sentiment labels were changed to a binary format for all datasets. For the translated datasets, noise removal was performed prior to tokenization, similar to the “Remove Noise” step performed for the baseline classifiers (see Section 3.2.1). 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 ob-

[11] https://huggingface.co/docs/transformers/main/tokenizer_summary#byte-level-bpe

[12] https://huggingface.co/docs/transformers/main/tokenizer_summary#wordpiece

Classifier	English	Google	Greynir
Support Vector Classifier	**89.68%**	**89.02%**	**88.15%**
Naive Bayes	85.79%	85.78%	85.16%
Logistic Regression	89.35%	88.74%	87.76%
RoBERTa	**94.90%**		
IceBERT		92.18%	90.74%
ELECTRA		**92.24%**	**92.16%**

Table 7: Accuracy of the baseline classifiers and the transformer models on the original English IMDb dataset (column 2) and on the translated datasets (columns 3 and 4).

servations indicated that transformer models yield better results when trained on more diverse corpora, 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. The default parameters can be seen in the Appendix.

For training the transformer models, we used the AdamW optimizer (Loshchilov and Hutter, 2019).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 were trained 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.

4. Results

In this section, we provide evaluation results, for the baseline classifiers, on the one hand, and the transformer models, on the other, for both translated and non-translated data.

4.1. Baseline Classifiers

The accuracy of each baseline classifier trained on the English dataset and the machine-translated datasets are shown in Table 7. The best-performing baseline classifier for the translated Icelandic datasets is the Support Vector Classifier (SVC), which achieved an accuracy of 89.02% on the data translated by Google Translate[13]. Thus, the best Icelandic SVC model is only 0.66% less accurate in determining the sentiment of IMDb movie reviews than the best English model.

Translation Service	Accuracy	SD
GreynirTranslate	90.9%	1.69
Google Translate	91.5%	1.36

Table 8: The average accuracy and standard deviation of the ELECTRA model, fine-tuned on data translated by either GreynirTranslate or Google Translate, when evaluated on original Icelandic movie reviews.

4.2. Transformer Models

The accuracy of the transformer models are shown in Table 7. The RoBERTa model obtains an accuracy of 94.9% on the original English IMDb dataset. For the translated Icelandic datasets, ELECTRA obtains the highest accuracy of 92.24% on data translated by Google Translate[13]. Thus, the best Icelandic transformer model is 2.66% less accurate than the English RoBERTa model.

4.3. Icelandic Reviews

We evaluated the best performing model, trained on translated data (i.e. ELECTRA), on movie reviews originally written in Icelandic. We ran the evaluation 10 times with 100 sampled reviews split evenly into 50 positive and 50 negative reviews, and averaged the accuracy. The results, shown in Table 8, show that ELECTRA fine-tuned on translations produced by GreynirTranslate and Google Translate obtained an accuracy of 90.9% and 91.5%, respectively.

5. Discussion

Our work outlines a methodology for developing ML models for sentiment analysis of Icelandic movie reviews by using machine-translated data for training. 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 Icelandic texts, that have been translated

[13] McNemar's test (McNemar, 1947) shows a statistically significant difference between the classifiers trained on data translated by Google Translate and data translated by GreynirTranslate.

from English, 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 2.66%. Additionally, we found support for the claim that models trained on the original English dataset would obtain the highest accuracy. Our evaluation shows that the RoBERTa model trained on English data performed the best of all the models, obtaining an accuracy of 94.9%.

We found evidence across all of the models in support of our second hypothesis, that models trained on data translated by Google Translate would obtain the highest accuracy. The most accurate baseline model was the Support Vector Classifier, trained using data translated by Google Translate, with an accuracy of 89.02%. The most accurate transformer model was ELECTRA, fine-tuned using data translated by Google Translate, with an accuracy of 92.24%. Comparatively, the RoBERTa model, which is fine-tuned on the original English data, achieved an accuracy of 94.9% – thus, the drop in accuracy is 2.66%.

The third and last hypothesis was that IceBERT (a RoBERTa model) would obtain the highest accuracy amongst the transformer models. We did not find support for this, since the Icelandic ELECTRA model obtained the highest accuracy on the translated data. This is an interesting result, because the the ELECTRA model is pre-trained on considerably less data than the IceBERT model. Both models use the IGC for pre-training, but, in addition, IceBERT uses web data. Thus, the lack of web data as part of the pre-training data for the ELECTRA model does not seem to make a difference for this sentiment analysis task.

We also note that the accuracy is similar when evaluating the model on Icelandic non-translated data. ELECTRA, fine-tuned using data translated by GreynirTranslate, achieved an average accuracy of 90.9% and, when fine-tuned using data translated by Google Translate, the same model obtained an average accuracy of 91.5%.

We observed that the translated texts from both GreynirTranslate and Google Translate are most often syntactically correctly, and that the semantic meaning of the text in both cases transfers when sentiment analysis is carried out on the translations.

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

6. Conclusion

Our study demonstrates the effectiveness of leveraging machine-translated data for sentiment classification in Icelandic, where no such dataset previously existed. Through the automatic translation of 50,000 English IMDb reviews into Icelandic using two translation services, we evaluated the retention of sentiment in the target language and assessed the accuracy of sentiment classifiers on non-translated Icelandic text. Our analysis compared three types of baseline classifiers with three pre-trained transformer-based models (RoBERTa, IceBERT, and ELECTRA) on both original English texts and translated texts. Our findings reveal that transformer models outperform baseline classifiers across all datasets, indicating their superiority in sentiment classification tasks. Additionally, we showed that transformer models trained on data translated from English reviews effectively classify sentiment in native Icelandic movie reviews. These findings are promising for the task of sentiment analysis in Icelandic and may generalize to other (low-resource) languages for which a large corpus of sentiment data is not available.

In future work, we would like explore the feasibility of employing our methodology for various other classification tasks in Icelandic, such as emotion detection, spam detection, and topic categorization. We are also interested in the effectiveness of data augmentation methods for low-resource languages to increase available dataset for NLP tasks, such as text classification, e.g., back-translation, synonym replacement, or text generation.

7. Limitations

In our research, several constraints were noted. The first concerns time constraints and computational resources required. Training transformer models can be time-consuming and resource-intensive, but this is contingent on the dataset provided for the model. Second, our methodology may not generalize to other domains beyond sentiment classification on movie reviews. Other domains and tasks may require bespoke approaches to data collection and processing, as well as modeling methods. Furthermore, while Transformer models are powerful, they are often seen as “black boxes”. The lack of interpretability can be a significant limitation, especially when trying to understand the factors contributing to the model’s classification of new reviews or when errors need to be diagnosed.

8. Bibliographical References

K. Amulya, S. B. Swathi, P. Kamakshi, and Dr. Y. Bhavani. 2022. Sentiment Analysis on IMDB Movie Reviews using Machine Learning and Deep Learning Algorithms. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 814–819.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Jón Friðrik Daðason and Hrafn Loftsson. 2022. Pre-training and Evaluating Transformer-based Language Models for Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7386–7391, Marseille, France. European Language Resources Association.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ayanabha Ghosh. 2022. 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)*, 1:289–294.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland. Linköping University Electronic Press.

Borislav Kapukaranov and Preslav Nakov. 2015. Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository*, arXiv:1907.11692.

Hrafn Loftsson and Robert Östling. 2013. 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)*, pages 105–119, Oslo, Norway. Linköping University Electronic Press, Sweden.

Pintu Lohar, Maja Popović, and Andy Way. 2019. Building English-to-Serbian Machine Translation System for IMDb Movie Reviews. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 105–113, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011b. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Alberto Poncelas, Pintu Lohar, James Hadley, and Andy Way. 2020. 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), pages 78–88, Virtual. Association for Machine Translation in the Americas.

Saeed Mian Qaisar. 2020. Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pages 1–4.

Francisco Javier Ramírez-Tinoco, Giner Alor-Hernández, José Luis Sánchez-Cervantes, Beatriz Alejandra Olivares-Zepahua, and Lisbeth Rodríguez-Mazahua. 2018. A Brief Review on the Use of Sentiment Analysis Approaches in Social Networks. In *Trends and Applications in Software Engineering. CIMPS 2017. Advances in Intelligent Systems and Computing, vol 688*. Springer.

Gayane Shalunts, Gerhard Backfried, and Nicolas Commeignes. 2016. The Impact of Machine Translation on Sentiment Analysis. In *The Fifth International Conference on Data Analytics*, pages 51–56, Venice, Italy.

Iyanuoluwa Shode, David Ifeoluwa Adelani, JIng Peng, and Anna Feldman. 2023. NollySenti: Leveraging Transfer Learning and Machine Translation for Nigerian Movie Sentiment Classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 986–998, Toronto, Canada. Association for Computational Linguistics.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

9. Language Resource References

Jón F. Daðason. 2017. *Nefnir: A lemmatizer for Icelandic text*. Github.

Jón F. Daðason. 2022. *Icelandic ELECTRA-base*. Hugging Face.

FacebookAI. 2019. *RoBERTa base*. Hugging Face.

Atli Jasonarson. 2018. *Icelandic Stop Words*. Github.

Hrafn Loftsson. 2009. *IceNLP*. Github.

Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher. 2011a. *IMDB Dataset of 50K Movie Reviews*. Distributed via Kaggle.

Miðeind. 2022. *IceBERT*. Hugging Face.

Vilhjálmur Þorsteinsson and Hulda Óladóttir and Sveinbjörn Þórðarson and Pétur Orri Ragnarsson and Haukur Páll Jónsson and Logi Eyjólfsson. 2022. *Tokenizer for Icelandic text (3.4.1) (2022-05-31)*. CLARIN-IS.

Snæbjarnarson, Vésteinn and Símonarson, Haukur Barri and Ragnarsson, Pétur Orri and Jónsson, Haukur Páll and Ingólfsdóttir, Svanhvít Lilja and Þorsteinsson, Vilhjálmur. 2021. *GreynirTranslate - mBART25 NMT models for Translations between Icelandic and English (1.0)*. CLARIN-IS.

10. Appendix

Classifier	Default parameters
Naive Bayes	alpha=1.0, fit_prior=True, class_prior=None
Support Vector Classifier	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 9: Parameters for the baseline classifiers.