UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science master thesis

Exploring the integration of RobBERT as a contextual embedder in the sentiment analysis of historical text

First examiner:

Pim Huijnen

Second examiner:

Antal van den Bosch

Candidate:

Sander Daniël Scheeper

June 29, 2024

Abstract

Understanding sentiment in historical texts offers valuable insights into public opinion on historical events. This study explores the use of a RobBERT-based language model for sentiment analysis on Dutch historical newspaper articles. Given the objectivity of these articles, sentiment classification poses unique challenges. We show that the RobBERT-based pipeline struggles with accurate sentiment classification, performing inadequately for reliable classification. Cohort analysis revealed a slight positive relationship between prediction confidence and accuracy, and error characterization found no significant differences between correct and incorrect classifications. These results highlight the limitations of using unoptimized models for historical text sentiment analysis. Enhancing dataset size, improving contextual understanding by creating a fill-mask paradigm, or incorporating human-in-the-loop methods may improve performance. This study underscores the need for adapted models to better analyze historical sentiments.

Contents

1 Introduction **3**

2 Data **7**

2.1 Data 7

2.2 Preprocessing 8

3 Methods **10**

3.1 Pipeline creation and evaluation 10

3.2 Cohort analysis 10

3.3 Error characterization 11

4 Results **12**

4.1 Pipeline evaluation 12

4.2 Cohort analysis 12

4.3 Error characterization 14

5 Discussion **17**

6 Appendix **20**

Bibliography **22**

1. Introduction

The discovery of the gas field under Groningen in 1959 was a huge boost for the economy of the Netherlands. Though, even before that, fossil fuels played a role in the Dutch society. Already in the nineteenth century, coal mines were opened in the southern province of Limburg. These same coal mines were closed in 1965, as other fossil fuels, like oil and gas, became more profitable than coal (Schot et al., 2000). These closings were finalized in 1975 when the last mine, Emma-Hendrik, was closed. In that same year, an oil-crisis swept over the Netherlands, when the Yom-Kippur war between Israel and the Arabic countries caused the oil prices to be raised significantly, and a boycott was installed against countries that supported Israel, including the Netherlands. All of these events were significant to the evolution of energy in the Netherlands, and the general public felt the impact of these events in their daily lives: Car-less Sundays, and much more focus on well-functioning energy policies (Schot et al., 2000).

Opinions about significant events and the evolution of energy is much easier to document today, platforms like X (previously Twitter) and Meta (previously Facebook) contain multitudes of opinions on fossil fuels. However, for pre-social media periods, platforms for sharing opinions were more scarce and less documented. One of the most documented sources of opinions and historical events are newspapers (Bingham, 2021). Especially in the 20th century, newspapers were a common form of information distribution. They established themselves as interpreters, and representatives, of popular opinion (Bingham, 2021). Many papers also claim "impartiality" or "independence", and while this might seem true, the way a paper frames a topic, through language or other means, impacts the way people react to the news (Bingham, 2021). This means that opinion and sentiment in newspapers is particularly difficult to extract, as overt language use becomes less important for reading the sentiment towards a topic. Even though the use of objective language is apparent, opinion is still subtly present in these newspaper articles.

The extraction of these opinions, or sometimes referred to as sentiments, can prove an extensive task, especially when done by human labelers. Methods therefore exist to alleviate the burden and speed up classification processes with domain-appropriate accuracy. Natural Language Processing (NLP) methods read the sentiment through several different methods:

First of all, lexicon- or linguistic-based approaches read the sentiment by looking at the sentiment scores in a pre-defined lexicon (Taboada et al., 2011). These sentiment scores often only classify between negative and positive, like VADER (Hutto & Gilbert, 2014) and AFINN (Nielsen, 2011), but other lexicons can classify a text into many different emotions, like EmoLex (Mohammad & Turney, 2010). Even though these methods work well for modern data, the meaning of a word changes over time, which can skew analysis results in cases where the time between the model development and the analyzed text are too far apart. For example, language between 1949 and 1968 experienced a rather rapid change (Juola, 2003). The meaning of words become different, and certain words can become archaic. This becomes especially important when doing text analysis on historical text.

Second of all, Machine learning classifiers like Support Vector Machines (SVMs) and Logistic Regression (LR) base their classification on pre-labeled training data. This has the advantage that lexicons do not have to be updated, and that only a small set of cases have to be labeled in order for the model to train itself. However, this method can suffer from labeler bias, which can occur when a person's own cognitive biases and subjectivity influence their choices in a labeling task (Brodley & Friedl, 1999). This can result in a model predicting on the cognitive biases of the labelers, rather than the true sentiment of the data. Regardless, these methods are more resistant to the changes of the meaning of words, and are therefore mores suited for historical text classification than lexicon based methods.

In NLP, text representation is commonly achieved through vectorization techniques, traditionally implemented using models like Word2Vec or

GloVe. However, bi-directional transformer encoders, exemplified by models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), can also create contextual embeddings that perform well in use-cases like sentiment analysis, topic classification, or spam detection. In many cases, neural network classification heads are put on top of these encoders to completely move away from traditional machine learning methods.

However, the use of machine learning in a pipeline involving transformers should not be underestimated. They often require less computational power than transformer models, especially during training. Their results are also far more interpretable than those of transformer models, since machine learning methods use simpler mathematical algorithms for classification, rather than deep neural networks with many hidden layers.

The interplay between machine learning and encoder transformers has been studied in modern examples. For example, Fahim et al. (2023) demonstrated that in sarcasm detection, contextual embeddings using BERT far outperformed traditional vectorization methods (Word2Vec, GloVe), and a bidirectional recurrent neural network model called BiGRU. In another study by Mollah et al. (2024), RoBERTa was used as contextual embeddings for XGBoost, which outperformed their earlier findings with traditional vectorization techniques. This combination of XGBoost and RoBERTa even outperformed a RoBERTa classifier with a classification head on top, which was because computational resources limited the training of the RoBERTa model.

This work shows that contextual embeddings generated by bidirectional transformers outperform traditional classification methods, while maintaining computational efficiency. However, a critical limitation of this research is the focus on modern data. Evaluating these methods on data with characteristics less familiar to the model can bring to light shortcomings and identify challenging characteristics of the text itself. Identifying these shortcomings is important for understanding the linguistic changes across historical periods. It would not only reveal the model's limitations with this kind of data, but also offer valuable insights into the changes in language across history. This knowledge can be leveraged to develop targeted training strategies for these use cases, improving the model's ability to process and analyze historical texts effectively.

Therefore, this study aims to investigate the effectiveness of a Dutch BERT-based model for generating contextual embeddings within a sentiment classification machine learning pipeline. Specifically, we will employ RobBERT-2023-dutch-large (Delobelle & Remy, 2023), a Dutch RoBERTa-based model developed by KU Leuven, UGent, and TU Berlin. The research will focus on three key aspects: Firstly, A pipeline will be created that leverages RobBERT as a contextual embedder to perform sentiment classification. Secondly, the outcome of this pipeline will include likelihood scores, which will allow us to identify cases with low confidence scores which require human labelling. Lastly, an analysis of incorrectly classified paragraphs will be conducted, which will identify patterns and characteristics within these texts that contribute to model shortcomings.

2. Data

2.1 Data

The data consists of a collection of 2144 newspaper articles, obtained from the KB, the Dutch national library. Newspapers were digitized using OCR, and their features classified into distinct categories: advertisements, articles, and images. For this research paper, only articles were used. Paragraphs that were longer than 400 words were split into separate paragraphs, and paragraphs that consisted of too little words were discarded, although the exact cut-off value is unknown.

The data that resulted from this wrangling is grouped per decade, ranging from the 1960s to the 1990s, resulting in a total of 4 groups. For each year, these datasets were then grouped per fuel type (oil, coal, or gas), which resulted in a total number of 12 datasets, each containing the data for the corresponding year + fuel type.

Next, the sentiment of these paragraphs were labeled by three labelers. Though originally, two individuals were tasked with labeling the paragraphs. These labels ranged from 0 to 2 (0 being negative, 1 being neutral, and 2 being positive). The judgements of these labelers were weighted in the following way: If two labelers had opinions with zero steps in-between, the extreme opinion was always chosen. If the two labelers had the same opinion, the label remained the same. If the two original labelers had diverging opinions, the example was labeled again by the third labeler, from which the most frequent label was preferred.

2.2 Preprocessing

For this research paper, additional preprocessing was needed. During exploration, it was noticed that one article could belong to one or more fuel types. This resulted in an article being able to be duplicated, with a label for each fuel type. In a machine learning model, duplicates can lead to confusing results, most often overfitting or underfitting, depending on if the label is also the same in the duplicated instance. To prevent these problems. The datasets were first concatenated into one singular dataset containing all data. Each dataset contained the text, label, fuel type, and decade of the article. This dataset was then split into a train and test set. With the training data containing 80% of the data, and test data containing 20% of the original data. The train and test sets were then split for each fuel type, resulting in a total of three datasets. Each dataset contained an average of 551 articles, with separate training and test sets. Label distribution was analyzed to check whether sampling was needed. As is visible in figure 2.1, the positive label was the majority class in all datasets. Multiple sampling methods were considered, but random under-sampling was chosen to keep as much interpretability of the data as possible. The under-sampling was only performed on the training dataset, as to keep the original distribution for testing purposes. The resulting distributions can be seen in figure 2.2.

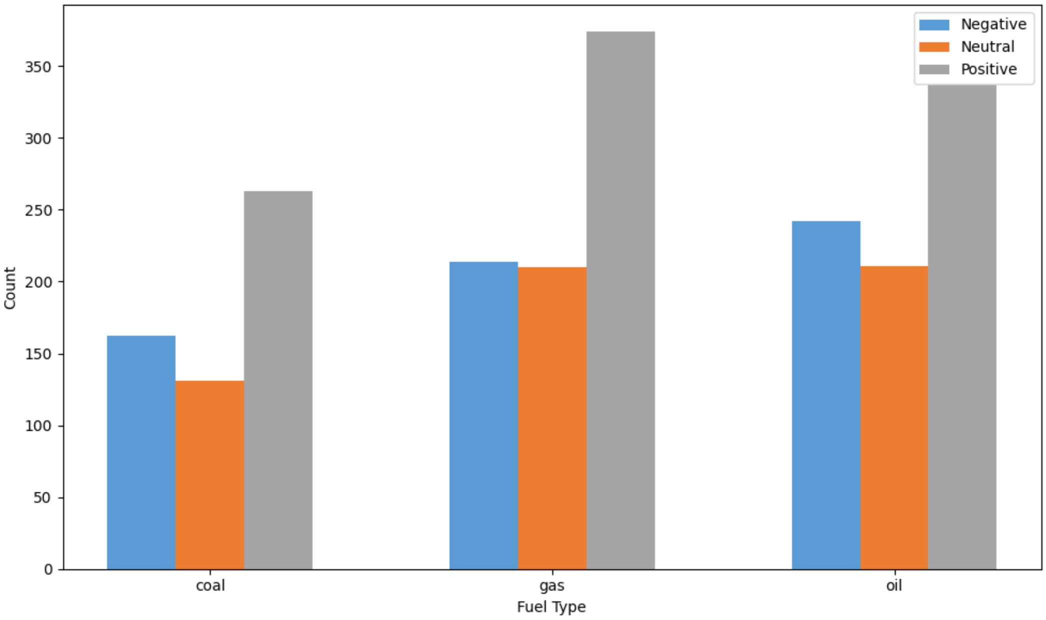

Figure 2.1: Label distribution within each fuel type dataset

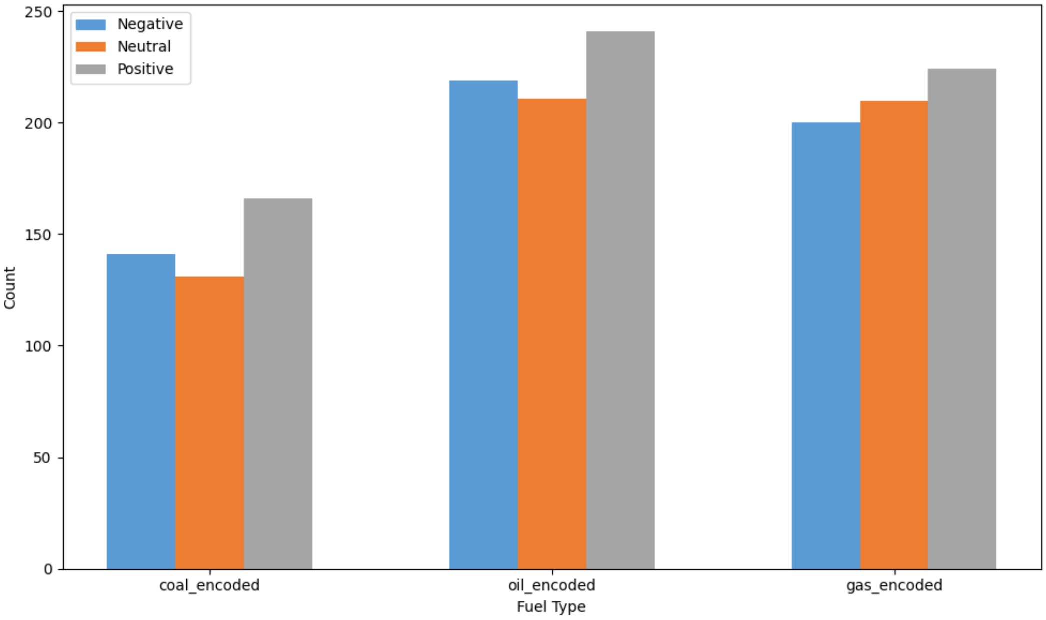

Figure 2.2: Label distribution after random under-sampling

3. Methods

3.1 Pipeline creation and evaluation

An untrained RobBERT model was initially loaded, and a feature extractor was constructed for it. This feature extractor processed batches of input data through the model to extract the last hidden states, which then returned the vector for the classification token ([CLS]) . This token represents the whole input as one vector (Wu et al., 2023), resulting in one vector per article. These vectors were used in a logistic regression classifier, utilizing an L2 penalty and a Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) solver. This classifier was trained on the extracted features, and its performance was assessed using confusion matrices, along with accuracy and f1 score as the most important evaluation metrics.

The pipeline was compared against a logistic regression classifier using 160-dimensional vectors generated by a dutch Word2vec model as a baseline (Tulkens et al., 2016). Text preprocessing for this pipeline involved lowercasing, removing punctuation, removing numbers, tokenizing the text using nltk (Loper & Bird, 2002), and then lemmatization using spaCy (Honnibal & Montani, 2017), which are standard preprocessing steps when using a complete machine learning pipeline.

3.2 Cohort analysis

An analysis of prediction probabilities was conducted after classification. Cohorts based on the probability of the predicted label were created, specifically: [′<0.5′, ′0.5-0.6′, ′0.6-0.7′, ′0.7-0.8′, ′0.8-0.9′, ′0.9-1.0′]. This stratification allowed for the evaluation with accuracy and f1 score as metrics within each probability range.

3.3 Error characterization

Finally, an error characterization was performed, examining the nature and distribution of misclassifications. Firstly, the distribution of correct and incorrectly classified articles were plotted for the decade to which the article belonged. In addition to this, several more advanced methods were used to compare the correct and incorrect articles.

3.3.1 Part-of-speech (POS) tagging

Part-of-speech (POS) tagging is the process of assigning grammatical categories to individual words within sentences. This method was used on a paragraph level, counting the total number of categories per paragraph. Using the SpaCy library (Honnibal & Montani, 2017) and their nl_core_news_sm model, the mean POS counts per article were calculated. This approach aimed to identify significant grammatical differences between articles that were correctly and incorrectly classified by the pipeline.

3.3.2 Syntactic dependency analysis

Syntactic dependency analysis examines the relationships between words in a sentence. It tries to determine the words that rely on other words to convey meaning. Here, the SpaCy library (Honnibal & Montani, 2017) and its nl_core_news_sm model were used again. The aim of this was to determine whether misclassifications could be attributed to difficulties with complex syntactic structures.

4. Results

4.1 Pipeline evaluation

The pipeline was evaluated using a confusion matrix and by comparing accuracy and f1 scores to a pipeline with Word2vec vectorization. Predictions on the three datasets (coal, gas, oil) were aggregated and treated as one while measuring performance. The confusion matrix for the logistic regression displayed poor performance, as seen in figure 4.1, with many positive articles being labeled as negative. The number of true positives, negatives, and neutrals barely, and in some cases, does not exceed other false classifications. Especially the neutral class seems to under-perform when looking at the confusion matrix. Accuracy and f1-score were used as metrics to compare the RobBERT contextual embedding to a baseline Word2vec vectorizer, which can be seen in table 4.1. The RobBERT pipeline outperforms the Word2vec pipeline slightly in terms of accuracy and f1 score, and when looking at the confusion matrix in figure 4.2, it can be seen that a similar pattern of distribution is followed for the classifications.

	Accuracy	F1 score
Word2vec	0.368	0.356
RobBERT	0.422	0.426

Table 4.1: Accuracy and F1 scores for the two vectorization methods using logistic regression

4.2 Cohort analysis

Cohorts were created from probability scores for the chosen label of each prediction. The same evaluation metrics were used on these cohorts as on the entire logistic regression. The confusion matrices (Figure 4.3) for these cohorts show no significant difference in performance. When looking at

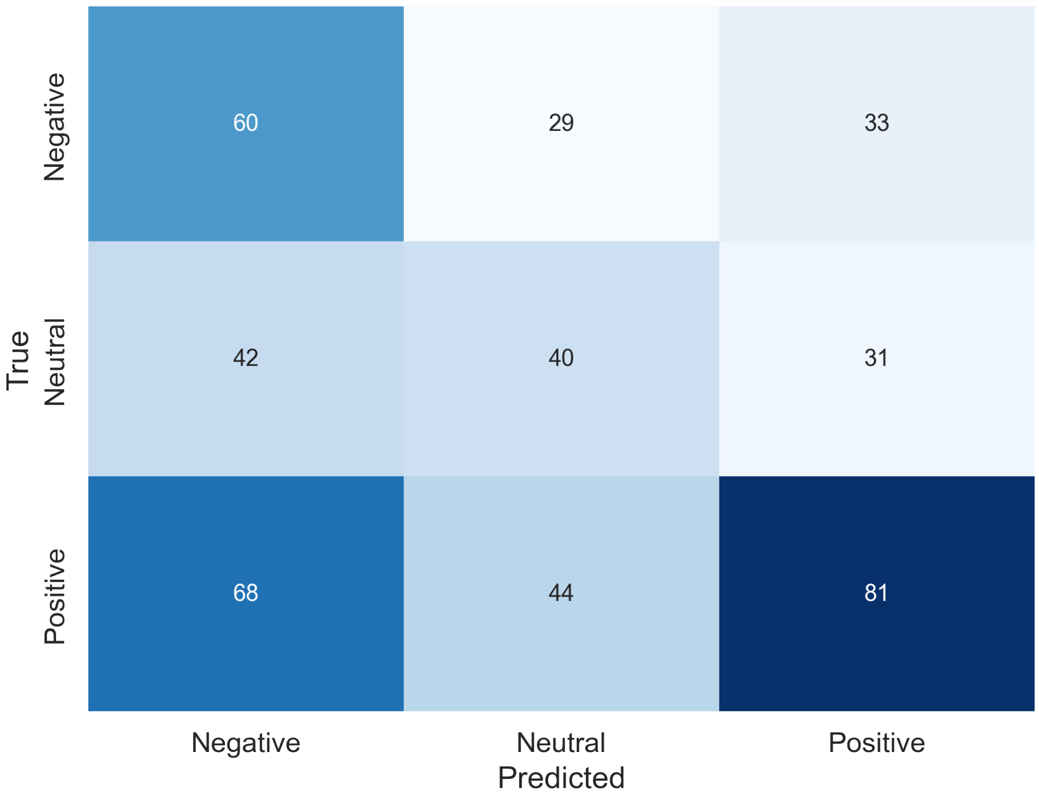

Figure 4.1: Confusion matrix of RobBERT pipeline in the aggregated dataset

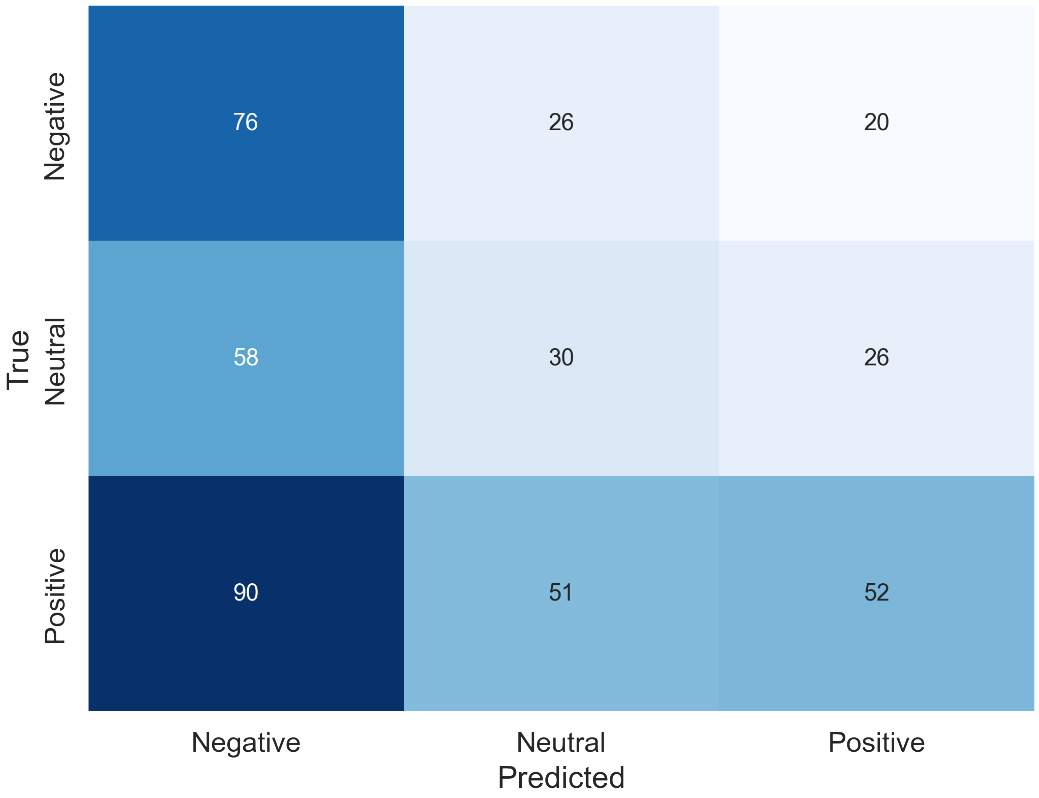

Figure 4.2: Confusion matrix of Word2Vec pipeline in the aggregated dataset

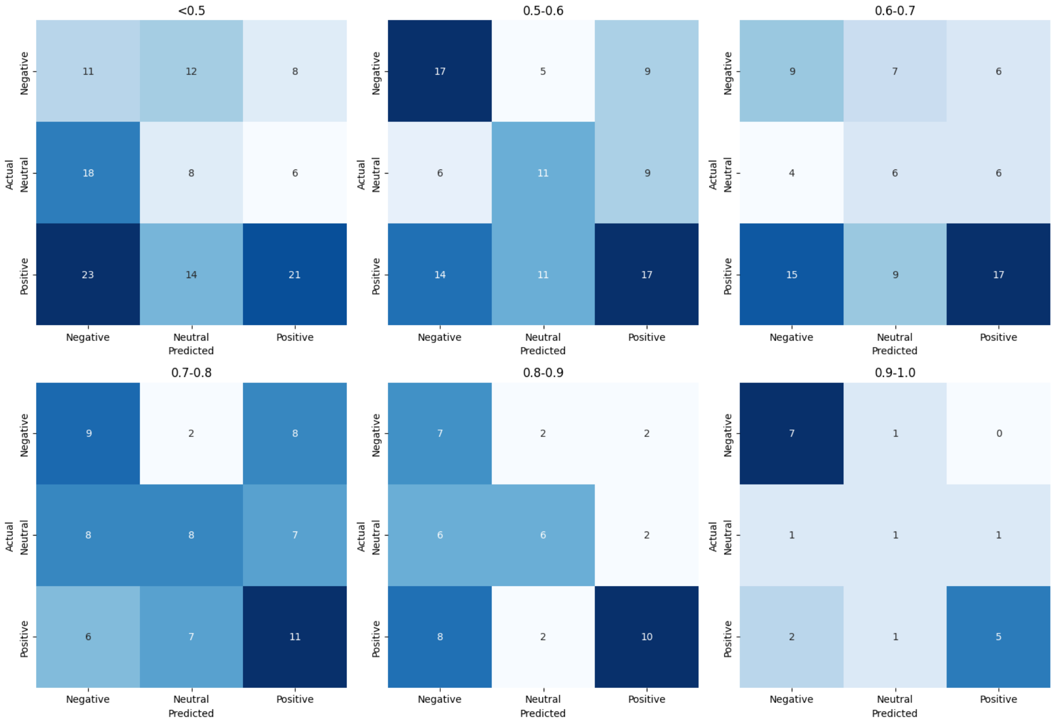

Figure 4.3: Confusion matrices of each probability cohort

metrics like accuracy and f1 score (Figure 4.4), we can see a slight increase in these metrics per cohort. With the highest cohort also having the highest performance in terms of metrics (accuracy = 0.684, f1 score = 0.681). It is important to note that the number of instances in this cohort was extremely low, with only 4.5% of instances being represented. Additionally, taking the confusion matrix for this cohort into account, it can be seen that classification for the neutral class still performs inadequately.

4.3 Error characterization

Correct and incorrect articles were first inspected per decade (Figure 4.5). No significant difference can be seen in the difference between the number of correct and incorrect articles in each decade.

4.3.1 POS tagging

POS tags showed significant overlap between correct and incorrect articles (Figure 4.6). Nouns and punctuation marks appear slightly more frequently

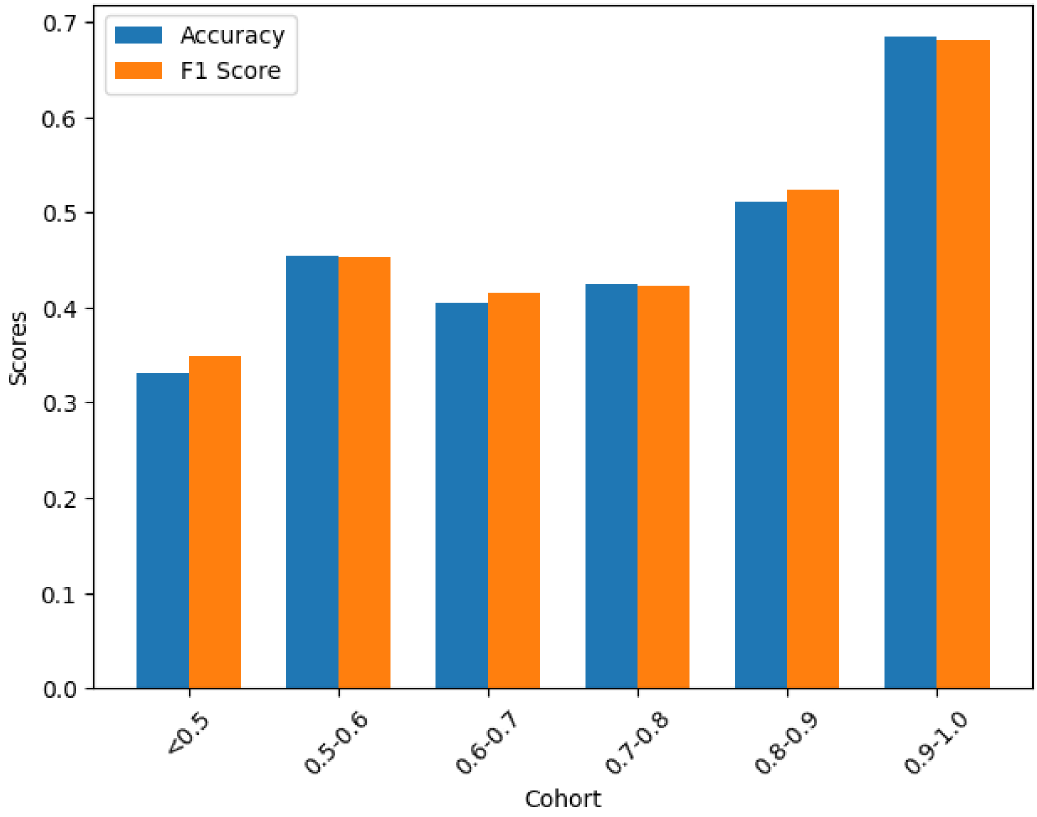

Figure 4.4: Accuracy and f1 scores of each probability cohort

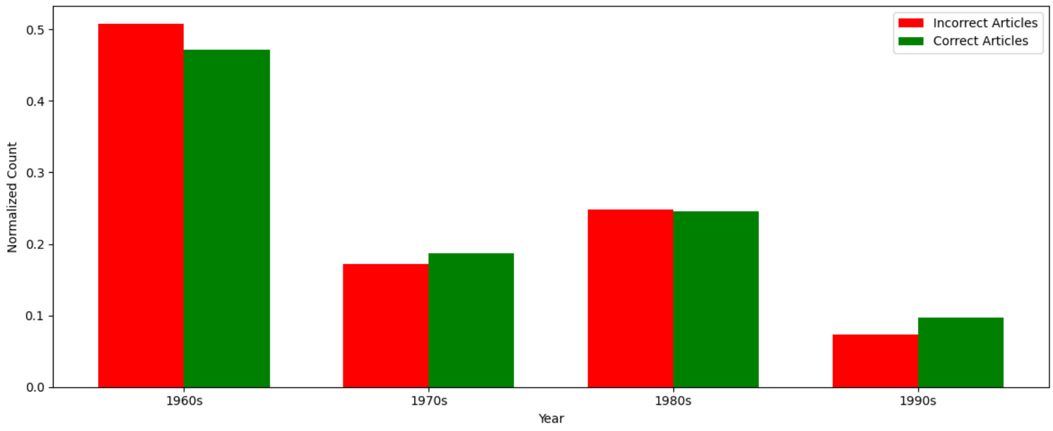

Figure 4.5: Normalized number of correct and incorrect articles per decade

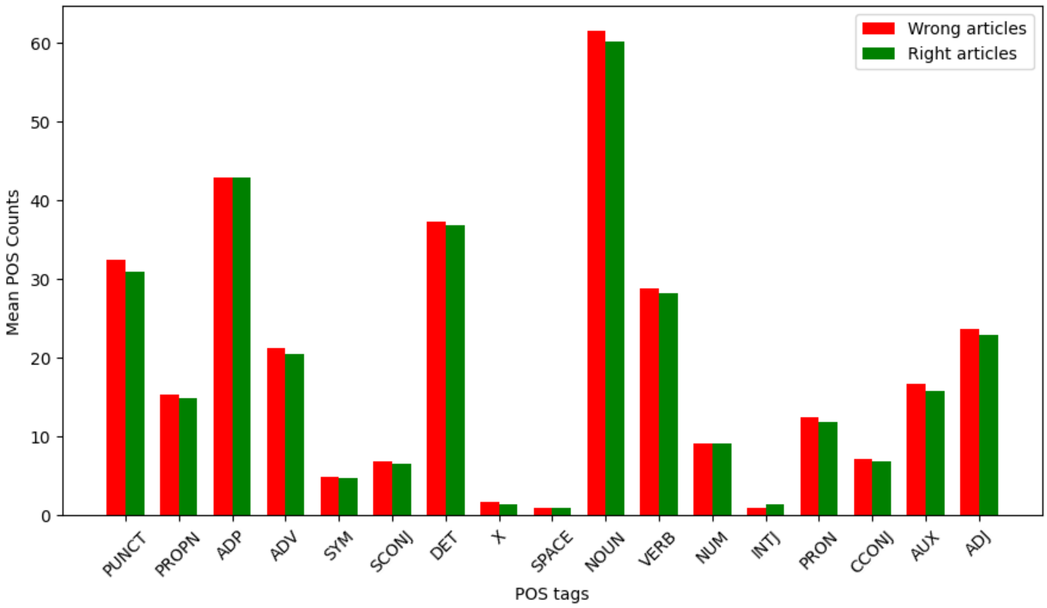

Figure 4.6: Mean POS counts per article category

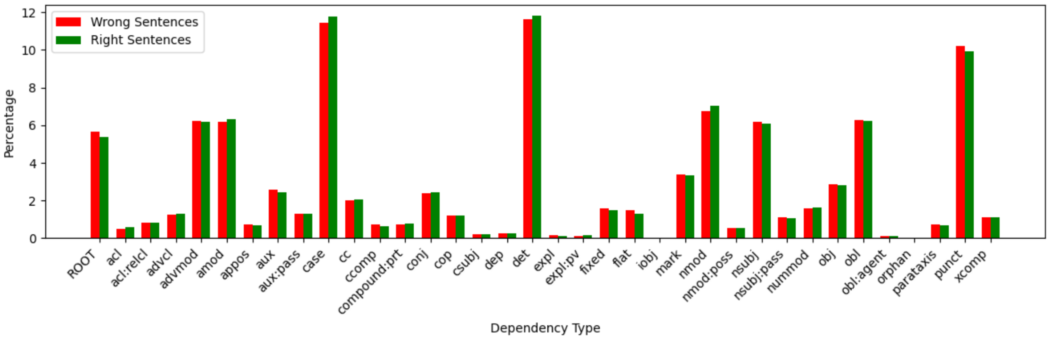

Figure 4.7: Dependency type percentages per article category

in incorrect articles, but this does not seem to be a significant difference.

4.3.2 Syntactic dependency analysis

Syntactic dependency analysis showed a significant overlap in dependency types between correct and incorrect articles (Figure 4.7). Determinants and case markers seem to appear slightly more frequently in correct articles, and root words and punctuation seem to appear slightly more frequently in incorrect articles. However, these differences do not seem significant.

5. Discussion

The central question of this research paper was to examine the viability of a RobBERT-based LLM as a contextual embedder within a machine learning pipeline for sentiment analysis on Dutch historical newspaper articles. This involved using standard evaluation metrics to measure performance and conducting error characterization to further assess the differences between correct and incorrect classifications made by the model. The goal of this process was to gain a deeper understanding of how a language model like RobBERT processes historical language and to assess the usability of this pipeline for researchers in the field of history.

A confusion matrix of the model's predictions showed that the number of correct classifications rarely, and in some cases, did not exceed the number of false classifications. It can be seen that the neutral class had the worst performance by far, with there being more neutral articles predicted as negative than neutral. Additionally, evaluation metrics showed that the RobBERT pipeline performed slightly better than the pipeline involving Word2vec. Regardless, accuracy and F1 score were poor, and suggested that this pipeline is still insufficient for effective classification.

During cohort analysis, an increase of accuracy and f1 score was expected with each cohort. The goal was to determine if filtering based on probability could provide results with only the most accurate predictions. A slight increase per cohort was found with some irregularities. The highest metrics were found in the highest probability cohort with an accuracy of 0.684, and an f1 score of 0.681. These results are promising, but given that this cohort only represents only 4.5% of the test data, and that the neutral class is also still underrepresented, these results are still not sufficient for the intended purpose of the pipeline.

In error characterization, I aimed to identify characteristics of the articles that could attribute to their correct or incorrect classification. It was expected that there would be differences per decade, considering that more recent decades could more modern language. However, decade-by-decade plots showed no significant differences. Similarly, POS tagging and syntactic dependency analysis also revealed no differences in these metrics between correct and incorrect articles. These results propose that differences in classification accuracy do not stem from differences in sentence structure, word types, or decade of the article.

The model's insufficient performance can be attributed to several factors. Firstly, as discussed previously, historical newspaper articles are considered more objective and less overtly subjective than modern sentiment classification data from social media. As the model relies on clear distinctions in the vector space between articles to make its classification, this lack of overt sentiment makes it challenging for the model to correctly discern articles. Secondly, the length of these articles means that the number sentimental words per non-sentimental words might be very small, diluting the model's ability to detect sentiment accurately. Additionally, there is the dataset's limited size, comprising only 2144 articles. After splitting this dataset into three fuel types, and undersampling the training data, the resulting datasets only had 551 articles on average. This might have caused the model to not be able to learn effectively. Additionally, the addition of a neutral class could also be considered a limitation, since it was seen that the neutral class had the worst performance as seen in the confusion matrix. Removing this neutral class, and keeping the classifications to positive and negative might allow the model to make more clear-cut choices, even though this might result in a worse representation of the data. Lastly, some articles were duplicated, where an article could belong to two or three fuel types at once. An attempt to account for this was made by splitting the dataset per fuel type. This type of splitting impacted the logistic regression training due to a significant decrease in the training data size but did not affect the vectorization method.

Possible next steps for research include the following. Increasing the size of the dataset could drastically help model learning and performance. Additionally, the models used in this paper could also be optimized for this type of data. For example, RobBERT could be trained on historical newspapers by creating a fill-mask task that allows the model to generate better contextual embeddings. This kind of task can easily be done on unlabeled data, removing the limitation of limited data resources. These improved contextual embeddings could then be used in a hybrid pipeline. An approach with a generative transformer can also be explored, where context is given within a given prompt, along with the article, to explain the differences in language from modern text. This might already provide some additional context that straightforward classification methods lack. Lastly, a human-in-the-loop approach could be designed, where domain experts can assist in labeling data or correcting model outputs, therefore improving model performance and reliability.

Besides these next steps, an interesting avenue could also be explored in labeler bias. In this case, since news articles are inherently trying to be objective, labelers might attribute their own opinions and emotions to the pieces of text more often than when a clear sentiment in the text is given. This increases bias in the training data, possibly also resulting in worse classification performance, since labelers may have diverging opinions.

While the current study has highlighted significant challenges in detecting sentiment in historical newspaper articles using a machine learning pipeline, it also opens avenues for future research. By addressing the identified limitations and exploring the proposed next steps, there is potential to develop a more effective sentiment analysis pipeline, and advancing the capabilities of historical natural language processing.

6. Appendix

Preprocessing and model pipeline code is available at github.com/Ketskapow/ADS-Thesis-Project

Bibliography

Bingham, A. (2021). Newspapers. *Bloomsbury History: Theory and Method Articles*. https://doi.org/10.5040/9781350970892.088

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, *11*, 131–167. https://doi.org/10.1613/jair.606

Delobelle, P., & Remy, F. (2023). Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion. *Antwerp, Belgium*. https://clin33.uantwerpen.be/abstract/robbert-2023-keeping-dutch-language-models-up-to-date-at-a-lower-cost-thanks-to-model-conversion/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fahim, K. M. H., Moontaha, M., Rahman, M., Rhythm, E. R., & Rasel, A. A. (2023). Comparative analysis of traditional and contextual embedding for bangla sarcasm detection in natural language processing. *2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 293–299. https://doi.org/10.1109/COMNETSAT59769.2023.10420673

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* [To appear].

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, *37*, 77–96. https://doi.org/10.1023/A:1021839220474

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. https://doi.org/10.48550/ARXIV.CS/0205028

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In D. Inkpen & C. Strapparava (Eds.), *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26–34). Association for Computational Linguistics. https://aclanthology.org/W10-0204

Mollah, M. A. R., Kabir, M. M. J., Reza, M. S., & Kabir, M. (2024). Adapting contextual embedding to identify sentiment of e-commerce consumer reviews with addressing class imbalance issues. *2024 International Conference on Advances in Computing, Communication, Electrical,* *and Smart Systems (iCACCESS)*, 1–6. https://api.semanticscholar.org/CorpusID:269316003

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. https://arxiv.org/abs/1103.2903

Schot, J., Lintsen, H., Rip, A., & de la Bruhèze, A. (2000). *Techniek in nederland in de twintigste eeuw, delfstoffen, energie, chemie*. Lecturis BV.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Tulkens, S., Emmery, C., & Daelemans, W. (2016). Evaluating unsupervised dutch word embeddings as a linguistic resource. In N. C. (Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*. European Language Resources Association (ELRA).

Wu, L., Zhang, W., Jiang, T., Yang, W., Jin, X., & Zeng, W. (2023). [cls] token is all you need for zero-shot semantic segmentation.