Classification and analysis of articles from different Estonian news portals

Hanna Britt Parman (Computer Science, MSc, Institute of Computer Science),
Kristiina Keps (Computer Science, MSc, Institute of Computer Science),
Anna Laaneväli (Continuing education learner, Faculty of Science and Technology)

UNIVERSITY OF TARTU
Institute of Computer Science

IT Akadeemia

Abstract

In this project we collected articles from seven different Estonian news portals, did some statistical analysis on the data and created models to classify the articles to different news portals. We achieved the best results with a CNN that achieved 85% accuracy on the test set.

Introduction

The aim of this project was to gather data from Estonian news portals, analyse it and create models that would assign a news portal to a given article. The news portals used in this project were ERR, Postimees, Elu24, Eesti Päevaleht, Telegram, Uued Uudised, Õhtuleht. Web scrapers were created to collect the available data from all these portals. Neural networks CNN and EstBERT were used classifying as well as other machine learning methods like naive Bayes, logistic regression, K-nearest neighbors, least-squares support-vector machine and random forests. The CNN achieved the best accuracy of 85%.

Data gathering

From each portal the publishing date, headline, topic, content and URL of articles were scraped (with the exception of no publishing date for the news from the portal of Uued Uudised).
The news portals were built very differently so each of them required individual handling. For Postimees, Elu24, Uued Uudised, Eesti Päevaleht, ERR and Telegram it was possible to gather all the articles, but Õhtuleht set some limitations on the data gathering process and the authors managed to get only around 240 articles from there. However, it was still included in some of the models and in statistical analysis. Most news portals also require a subscription to read whole articles and for Telegram the authors didn't have a subscription which means the articles are not at their full length

Statistical analysis

At first the analysis of the length of the titles and articles was done. It turns out that the headline statistics are quite similar among all portals but the length of the content is quite different.
From the portals that have a paywall, Telegram seems to offer the most free content, as we can see from the article content statistics in table 1.

Table 1. Average article headline and content statistics (only letter characters counted, rounded to the nearest integer)

Portal	Article headline		Article content	
	Avg. characters	Avg. words	Avg. characters	Avg. words
Postimees*	53	7	1 184	178
Päevaleht*	57	9	2 701	434
ERR	51	7	1 589	246
Elu24*	62	9	1 160	189
Telegram*	51	7	2 785	437
Uued Uudised	67	10	2 237	341
Õhtuleht	66	9	946	141

*Datasets only included publicly available article content

Statistical analysis

From analysing the topics of the articles collected it became apparent that how the topics are classified greatly varies between portals. For example in Päevaleht, all of the news were classified under seven general topics (e.g. "culture" or "sport") while in Postimees there were over 300 different topics, which also included some general topics but also more niche topics that only had <10 news articles (e.g. "holidays 2020" or "curling").

Python wordcloud package was used to make word clouds for both the original and lemmatised headline input. The cloud was initially overwhelmingly covered in stopwords, so they were removed to create a more informative output. A comparative example of word clouds made based on original and lemmatised input can be seen on figure 1.

Figure 1. Word clouds with maximally 50 words for the headlines from Päevaleht: original(left) and lemmatised(right)

Named entity recognition tools from the "estNLTK" python package were used to identify most named people from all news portals headlines.
Most named people in each portal were politicians mostly from Estonia but also some top international politicians.

CNN

Two different models were used for classification: a simple CNN and EstBERT. Both gave good results, but since EstBERT took much longer to train and needed more memory, the authors decided to put more effort on the CNN model. The architecture of the CNN model is on figure 2.

Figure 2. The architecture of the CNN model

The data was tokenized with a Stanza tokenizer and converted to vectors using a pretrained vector vocabulary. 50 000 of the most frequent words were used and others were encoded to unknown words.
The authors also tried to use stop-words, first with an existing list of Estonian stop-words and then with the most frequent words in the data. These however didn't improve the accuracy, so in the final model no stop-words were removed. Three different kinds of input were used as well: the titles, first 100 words of each article and the whole articles.

CNN results

The accuracies, F1-scores and losses of different models that were trained for 10 epochs on a dataset of around 6000 articles are presented on figure 3.

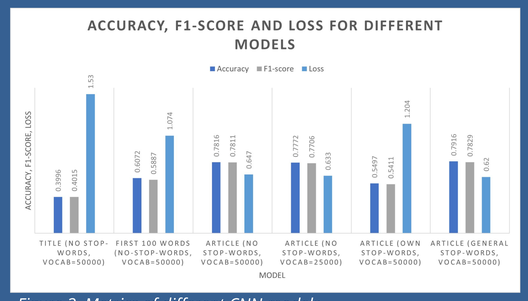

Figure 3. Metrics of different CNN models

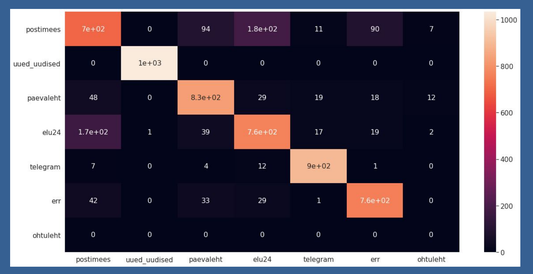

Figure 4. Confusion matrix of the best CNN model

Predicting the news portal based on just the title or the first 100 words didn't give good results. Removing the stop-words collected by us also lowered the accuracy. Using the general stop- words gave the same result sas using no stop- words. Changing the vocabulary size from 25 000 to 50 000 didn't affect the accuracy much. In the end the model with no stop-words and vocabulary size 50 000 was used for further training.
The final model was trained for 10 epochs on a training set of around 50 000 articles and it achieved accuracy of 85% on the test set. The confusion matrix of that model on the test set is presented on figure 4. Most of the articles were classified correctly with some mistakes (there were no samples from Õhtuleht in the test set).

Other machine learning models

In addition to CNN and EstBERT models we searched other ideas from the published papers on the similar topic. We found an article "Text classification: A least square support vector machine approach" by V. Mitra, C. Wang and S.Banerjee were document titles were classified between 6 different categories. In the article they had corpus of 91,229 words from University of Denver's Penrose Library catalogue and they had very high accuracy rate that caught our attention. Best accuracy was 99.9% with Least Squares Support Vector Machine (LS-SVM), followed by 92.7% with K-Nearest Neighbors (KNN) and by 89.3% with Naive Bayes (NB).

Results of other machine learning models

For our project we tried to predict news portal based on article headline. Each portal was classified headlines were tokenized. We used text data vectorization and document-train matrix, the stop words were not included as they didn't increase accuracy based on CNN and EstBERT models. Models we used for classification did not give nearly as good results as in article. Best accuracy of 59% gave NB, followed by Logistic Regression (LR) and Linear Support Vector Machine (LN-SVM) with 58% using full dataset of 23,5210 headlines. Using half of dataset didn't affect accuracy much- best was NB with 57%. By taking a random subset of 9000 articles from each portal (Õhtuleht was disregarded), accuracy plummeted to 41% in best case using NB. Best accuracy for KNN was 22% and for Random Forest it was 33%. In retrospective it wasn't worth to include these models although for KNN the results in article were promising.

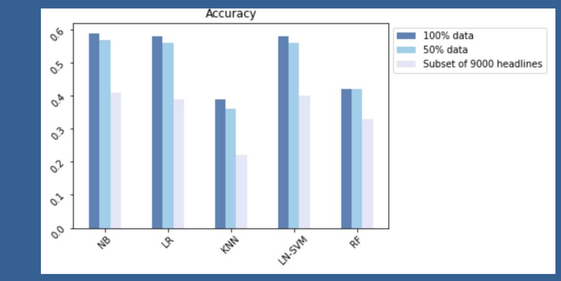

Figure 5. Accuracy of different machine learning models

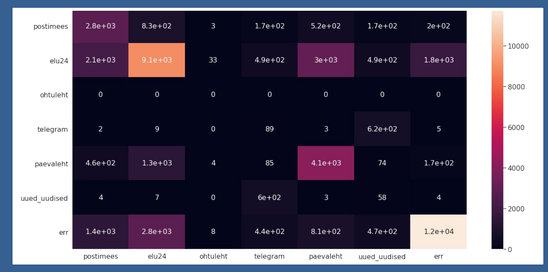

Figure 6. Confusion matrix of the best NB model

Conclusion and acknowledgements

The aim of this project was to see if it's possible for machine learning models and neural networks to distinguish articles from different news portals. Our CNN model achieved accuracy of 85% which implies that the characteristics of articles from different portals are in fact different enough to be classified correctly by a neural network.

The code for this project and a more detailed report is available in a GitHub repository:
https://github.com/kristiinakeps/est-news-portal-classification

This project was funded by the University of Tartu.