Classification and analysis of articles from different Estonian news portals

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

Abstract

In this project we collected articles from seven different Estonian news portals, did some statistical analysis of 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 Paevaleht, Telegram, Used Ududised, Onnuleht. 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, lead-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 Used Ududised). The news portals were built very differently so each of them required individual handling. For Postimees, Elu24, Used Ududised, Eesti Paevaleht, ERR and Telegram it was possible to gather all the articles, but Onnuleht 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 as some of the models and in statistical analysis. Most news portals also require a subscription to read whole articles and for Telegram the authors decided to subscribe which means the articles are not at their full length.

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

<table>
<thead>
<tr>
<th>Portal</th>
<th>Avg. characters</th>
<th>Avg. words</th>
<th>Avg. characters</th>
<th>Avg. words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postimees*</td>
<td>59</td>
<td>7</td>
<td>184</td>
<td>178</td>
</tr>
<tr>
<td>Eesti Paevaleht</td>
<td>57</td>
<td>9</td>
<td>2 701</td>
<td>434</td>
</tr>
<tr>
<td>ERR</td>
<td>51</td>
<td>7</td>
<td>1 589</td>
<td>246</td>
</tr>
<tr>
<td>Elu24*</td>
<td>62</td>
<td>9</td>
<td>1 690</td>
<td>189</td>
</tr>
<tr>
<td>Telegram*</td>
<td>51</td>
<td>7</td>
<td>2 785</td>
<td>437</td>
</tr>
<tr>
<td>Used Ududised</td>
<td>67</td>
<td>10</td>
<td>2 237</td>
<td>341</td>
</tr>
<tr>
<td>Onnuleht</td>
<td>66</td>
<td>9</td>
<td>946</td>
<td>141</td>
</tr>
</tbody>
</table>

*Datasets only included publicly available article content.

Statistical analysis

From analysing the topics of the articles collected it became apparent that the topics are classified by category between portals. For example in Paevaleht, all of the news were classified under seven general topics (e.g. "culture" or "sports") while in Postimees there were over 100 different topics, which also included some general topics but also more niche topics that only had <10 news articles (e.g. "holidays" 2020" or "curing").

Python wordcloud package was used to make word clouds for both the original and lemmatised headline input. The cloud was initially created by filtering the most used topics that 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.

Data from topic classification also shows that the headlines in Eesti Paevaleht and Telegram have a more informative content.

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.

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 high accuracy rate that caught our attention. Best accuracy was 90.9% 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 (LS-SVM) with 58% using full dataset of 23,5210 headlines. Using half of dataset did not affect accuracy much: best was NB with 59%. By taking a random subset of 9000 articles from each portal (Onnuleht was disregarded), accuracy plummeted to 41% in best case using NB. Best accuracy for KNN was 56% 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.

Conclusion and acknowledgements

This project was funded by the University of Tartu.

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

This project was funded by the University of Tartu.