UNIVERSITY OF COPENHAGEN

DEPARTMENT OF HUMANITIES - NorS

Creating a General Purpose Text Encoding Applied to Multiple Tasks

Niels Krog & Yasmin Shekari Goldbæk

Abstract

The aim of this thesis project is to apply a general-purpose encoding of Danish newspaper articles to authorship attribution, newspaper attribution, and headline generation with the aim of assessing the scalability of the encoding across task.

Introduction

The task and motivation

With the continual growth and availability of text online, the relevance of language understanding for text analysis has increased. Language understanding methods often rely on feature extraction processes, which can be done in *manual* or *automatic* manner. This has a wide range of applications, e.g. machine translation, information retrieval, and text summarization.

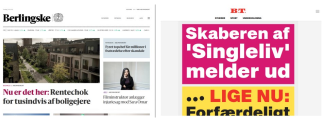
Manually extracted features have proven powerful in a wide range of tasks and are well studied. However, they are often tailored to a specific tasks and are time consuming to extract. On the other hand, the state of the art automatically extracted features have proven both powerful and versatile. In this project, we investigate the capabilities of both methods, given a range of different tasks.

Thanks to the company Ankiro, we got a hold of a novel dataset consisting of 808k Danish news articles from online sources. Based on this data, we settled on three tasks, leading us to the research question:

How can we create a general purpose feature encoding for authorship attribution, newspaper attribution and, headline generation?

To answer this question, we posed four subquestions:

1. Will manually extracted features, automatically extracted features of a combination of both generalize best?
2. Which of the manually extracted features has the highest impact on the results?
3. How do newspapers relate to each other? Does our feature representation reflect real world similarities between newspapers?
4. How will a manual encoding perform for headline generation?



Berlingske

Nu er det her: Rentechok for tusindvis af boligejere

Fyret topchef får millioner i fratrædelse efter skandale

Filminstruktør anlægger injuriesag mod Sara Omar

B.T.

Skaberen af 'Singleliv' melder ud

... LIGE NU: Forfærdeligt

Newspaper differences

Visual differences between the mainstream newspaper, Berlingske and its tabloid counterpart, BT.

Related works

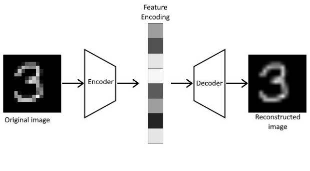
To design the methodology, we studied previous research on automatic text encoding methods and attempts at solving each of the aforementioned tasks in both a manual and automatic fashion. We found, that autoencoders could be used for automatically creating task-agnostic text encodings.

Authorship is without a doubt the most studied of the three tasks, dating back to the Middle-Ages and is the process of identifying the author behind a given text from a selection of known authors.

Newspaper attribution is the tasks of identifying from which newspaper a given article originates, given a known list of newspapers. The literature on this tasks is sparse, further motivating this project.

Headline generation can be seen as a sub-task within the field of text summarization and has been attempted using automatic methods, though no literature on using handcrafted features were found.

Finally, we investigated which types of manual features to use. *Content features* should be useful for newspaper identification, while *stylometric features* best for authorship attribution.



Autoencoders

Autoencoders can learn feature representations of data, here an image. The basic autoencoder works by having two neural networks: and encoder and decoder with a data bottleneck inbetween.

Methodology

Focus and expectations

The primary focus of this project is to explore the use of and scalability of a general-purpose encoding, with two tasks being classification tasks and one text generation task. Therefore, we have chosen to use the established classification models, *random forest and logistic regression*, instead of pursuing the best performing models to maximise the metrics. For the headline generation we used a pre-made LSTM autoencoder, validated on a similar task.

We expected the two classification tasks to be successful, especially because authorship attribution is such an established field of research. It seemed most likely, that combining manual and automatic features would yield the best result, as they are likely to complement each other. However, we did not expect the generated headlines to be impressive, mainly because it had not been attempted using manual features.

Cleaning up the data

The dataset consisted of *808,066* online news articles written by *66,994* authors across *268* domains. As the data is real-world data, it is far from being pristine. We performed a series of actions to clean up or salvage the entries:

- Removing duplicates, missing, and malformed entries
- Removing entries with multiple authors
- Removing entries with domain / author occurring only once
- Removing entries with foreign domain and duplicate headers
- Scraping articles

After the final preprocessing step, the dataset consisted of 301k articles. In the last step, a web scraper was implemented, updating some articles from certain domains.

Furthermore, we made an attempt at limiting data leakage in the article bodies. For example, if the author or newspaper name is present in the article body, it will compromise the classification tasks. Finally, we stripped the article bodies from junk text such as HTML tags or strings unique to a domain. The process of finding junk strings was stochastic and repeated twice.

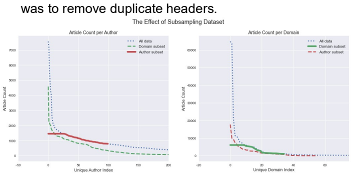
til at
menu søg
for at
log ind
altinget dk
læs mere
social sundhed
sundhed sverige
kontakt
udvikling ældre
og klima
få dages
fødevarer kommunal
cookie og
energi og
uddannelse udvikling
sverige transport
kommunal kultur
tde bedste
miljø social
forskning forsvar forsvar forsyning
sponsoreret indhold
civilsamfund civilsamfundets
eu forskning
kultur miljø

Word cloud of removed words

Many domains would contain repeating strings, such as ”log ind for at læse mere”. Such strings were removed.

Creating data subsets

After a thorough data insight and cleanup, the need for dividing the data into one subset for each of the tasks became apparent. The class distributions were skewed. Over- and underrepresented classes will disrupt the classification tasks. However, as classes are tied to each other, i.e. an article has both an author and domain, we made a subset for both classification tasks, independently limiting overrepresented classes and pruning underrepresented. As the headline generation is not a classification task, the only action for this subset was to remove duplicate headers.

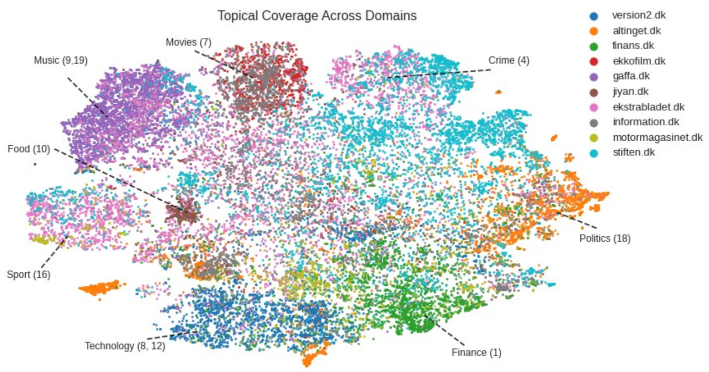


Data subset class distributions

The distributions of the subsets, according to articles per author (left) and domain (right). Zoomed in to show relevant subsamples

T-SNE topic coverage across domains

Each dot represents an article colored according to its domain, while the spatial position of each dot is determined by topic of the article.



Feature selection

The selection of features to include in the manual feature encodings were based on the related works. The features included were:

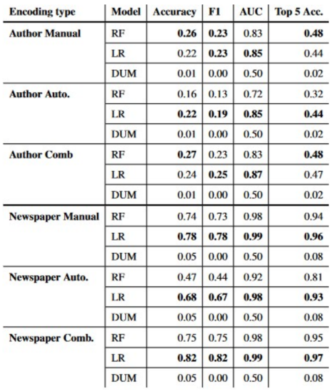
- Sentence / word statistics (count, length, etc)
- Hapax legomena count
- Lexical diversity measures
- Topic coverage
- Word n-grams
- Character n-grams
- POS n-grams
- Word Bi-skipgrams
- Function word frequency

Topic coverage was derived from an LDA model and the automatic features were derived from vector representation of the transformer model, BERT, pretrained on Danish text.

Evaluation and discussion

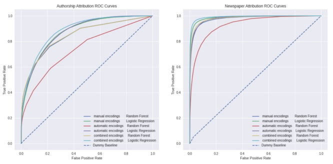
Classification results

The two classification tasks was evaluated using accuracy, F1 and area under curve (AUC) for the ROC curve. The classification models were compared against a dummy classifier as a baseline.



Encoding type	Model	Accuracy	F1	AUC	Top 5 Acc.
Author Manual	RF	**0.26**	**0.23**	0.83	**0.48**
	LR	0.22	**0.23**	**0.85**	0.44
	DUM	0.01	0.00	0.50	0.02
Author Auto.	RF	0.16	0.13	0.72	0.32
	LR	**0.22**	**0.19**	**0.85**	**0.44**
	DUM	0.01	0.00	0.50	0.02
Author Comb	RF	**0.27**	0.23	0.83	**0.48**
	LR	0.24	**0.25**	**0.87**	0.47
	DUM	0.01	0.00	0.50	0.02
Newspaper Manual	RF	0.74	0.73	0.98	0.94
	LR	**0.78**	**0.78**	**0.99**	**0.96**
	DUM	0.05	0.00	0.50	0.08
Newspaper Auto.	RF	0.47	0.44	0.92	0.81
	LR	**0.68**	**0.67**	**0.98**	**0.93**
	DUM	0.05	0.00	0.50	0.08
Newspaper Comb.	RF	0.75	0.75	0.98	0.95
	LR	**0.82**	**0.82**	**0.99**	**0.97**
	DUM	0.05	0.00	0.50	0.08

All classifications outperform the baseline and newspaper attribution has higher accuracy on all combinations, likely due to fewer classes than authorship attribution

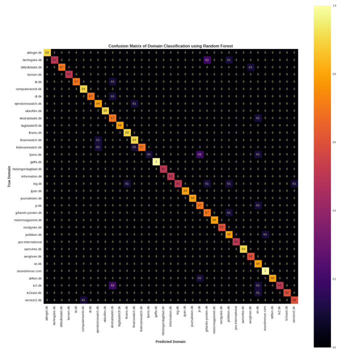


ROC curves for authorship (left) and newspaper attribution (right)

Newspaper attribution classification is better than authorship, Most likely beause of fewer classes.

Interpretation

Through feature importance insights, it was found, that the best performing features on both classification tasks were character bigrams and trigrams, and POS unigrams. While both classification tasks benefit from these features, the two tasks have different feature importances, indicating, that there is a clear difference between authorship and newspaper attribution. Topic coverage proved a useful feature for describing the characteristics for the newspapers (see top figure). Investigating the coverage confirmed our expectations in regards to how the newspapers related to each other.



Confusion matrix for newspaper classification

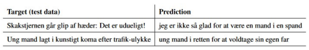
Confusion matrices were made for classification results. Here we see, that similar domains are confused with each other in some cases, but other cases there is a clear difference.

The results from the confusion matrices seemed suspiciously accurate, so we performed a series of tests to challenge our classification models. However, they passed the tests and were able to classify never seen before news articles. Though the models passed validation tests, it is a possibility, the models were biased, due to data leakages that were missed in the sheer amount of data.

Headline generation

While the headline generation worked when validating the model, it produced the same single headline for all articles. Although the loss kept decreasing, the process was too computationally heavy and we had to terminate it prematurely, as the results were unusable. For future works, experimenting with headline generation from manual feature encodings are encouraged.

The creation of a general purpose encoding for classification was successful, though not for headline generation.



Target (test data)	Prediction
Skakstjernen går glip af hæder: Det er udueligt!	jeg er ikke så glad for at være en mand i en spand
Ung mand lagt i kunstigt koma efter trafik-ulykke	ung mand i retten for at voldtage sin egen far

Humerous headlines

Generated headlines, which were either vulgar or oddly poetic.