UNIVERSITY OF HELSINKI
FACULTY OF ARTS
DEPARTMENT OF DIGITAL HUMANITIES

Master's Thesis

Supervised multi-class text classification for media research: augmenting BERT with topics and structural features

Ümit Naim Bedretdin

Master's Programme in Linguistic Diversity and Digital Humanities

Supervisor: Eetu Mäkelä 7.11.2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Humanistinen	Koulutusohjelma – Utbildningsprogram – Degree Programme Kielellisen diversiteetin ja digitaalisten ihmistieteiden maisteriohjelma	
Opintosuunta – Studieinriktning – Study Track Kieliteknologia		
Tekijä – Författare – Author Ümit Naim Bedretdin		
Työn nimi – Arbetets titel – Title Supervised multi-class text classification for media research: augmenting BERT with topics and structural features		
Työn laji – Arbetets art – Level Pro gradu	Aika – Datum – Month and year 11/2022	Sivumäärä– Sidoantal – Number of pages 35

Tiivistelmä – Referat – Abstract

Tämä työ esittelee ohjattuun koneoppimiseen perustuvan tekstiluokittelijan kehitysprosessin mediatutkimuksen näkökulmasta. Valittu lähestymistapa mahdollistaa mediatutkijan asiantuntijatiedon valjastamisen laaja-alaiseen laskennalliseen analyysiin ja suurten aineistojen käsittelyyn.

Työssä kehitetään neuroverkkopohjainen tekstiluokittelija, jonka avulla vertaillaan tekstistä erotettujen erilaisten luokittelupiirteiden kykyä mallintaa journalististen tekstien kehystystaktiikoita ja aihepiirejä. Kehitystyössä käytetyt aineistot on annotoitu osana kahta mediatutkimusprojektia. Näistä ensimmäisessä tutkitaan tapoja, joilla vastamedia MV-lehti uudelleenkehystää valtamedian artikkeleita. Siinä on aineistona 37 185 MV-lehden artikkelia, joista on eristetty kolme erilaista kehystystaktiikkaa (Toivanen et al. 2021), jotka luokittelijan on määrä tunnistaa tekstistä automaattisesti. Toisessa projektissa keskiössä on valtamedioissa käyty alkoholipolitiikkaa koskeva keskustelu, jota varten kerättiin 33 902 artikkelin aineisto Ylen, Iltalehden ja STT:n uutisista (Käynnissä oleva Vallan virrat -tutkimusprojekti). Luokittelijan tehtävänä on tunnistaa aineistosta artikkelit, jotka sisältävät keskustelua alkoholipolitiikasta. Työn tarkoituksena on selvittää, mitkä tekstin piirteet soveltuvat parhaiten luokittelupiirteiksi kulloiseenkin tehtävään, ja mitkä niistä johtavat parhaaseen luokittelutarkkuuteen.

Luokittelupiirteinä käytetään BERT-kielimallista eristettyä virketason kontekstuaalista tietoa, artikkelin muotoiluun liittyviä ominaisuuksia, kuten lihavointeja ja html-koodia, ja aihemallinnuksen avulla tuotettuja artikkelikohtaisia aihejakaumia. Alustavat kokeet pelkästään kontekstuaalista tietoa hyödyntävällä luokittelijalla olivat lupaavia, mutta niidenkään tarkkuus ei yltänyt tarvittavalle tasolle. Oli siis tarpeen selvittää, paraneeko luokittelijan suorituskyky yhdistelemällä eri piirteitä. Hypoteesi on uskottava, sillä esimerkiksi BERT-pohjaiset upotukset koodaavat muutaman virkkeen pituisen sekvenssin lingvististä ja jakaumallista informaatiota, kun taas aihemalli sisältää laajempaa rakenteellista informaatiota. Nämä piirteet täydentäisivät toisiaan artikkelitason luokitustehtävässä. Yhdistelemällä tekstien kontekstuaalista informaatiota aihemallinnukseen on hiljattain saavutettu parannuksia erilaisissa tekstinluokittelutesteissä ja sovelluksissa (Peinelt et al. 2020, Glazkova 2021). Yhdistämällä kontekstuaaliset piirteet aihemallin informaatioon päästään tässä työssä tosin vain marginaalisiin parannuksiin ja vain tietyissä ympäristöissä. Tästä huolimatta kehitetty luokittelija suoriutuu monesta luokittelutehtävästä paremmin, kuin pelkästään kontekstuaalisia piirteitä hyödyntävä luokittelija. Lisäksi löydetään potentiaalisia kehityskohteita, joilla voitaisiin päästä edelleen parempaan luokittelutarkkuuteen.

Kokeiden perusteella kehysanalyysiin perustuva automaattinen luokittelu neuroverkkojen avulla on mahdollista, mutta luokittelijoiden tarkkuudessa ja tulkittavuudessa on vielä kehityksen varaa, eivätkä ne vielä ole tarpeeksi tarkkoja korkeaa varmuutta vaativiin johtopäätöksiin.

Avainsanat – Nyckelord – Keywords
kieliteknologia, laskennallinen kehysanalyysi, mediatutkimus, BERT, finBERT, neuroverkot

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Abstract

This thesis showcases a workflow in developing a modern machine learning based classifier to bridge the gap between qualitative and quantitative research in media studies. Due to the recent datafication of our social environment, there has been growing interest in combining qualitative and quantitative methodologies in media studies. Current machine learning methods make it possible to gain insights from large datasets that would be impractical to analyze with more traditional methods. Supervised document classification presents a good platform for combining specific domain knowledge and close reading with broader quantitative analysis.

In this thesis, several classification features are extracted from journalistic texts and they are used to model framings and topics that are of interest to media researchers. Neural methods are utilized to build a supervised document classifier that can leverage the extracted features. The datasets used in development have been annotated as part of two ongoing media research projects. The first one consists of 37 185 articles from the Finnish countermedia publication MV-lehti and has been annotated into three categories based on a frame analysis of Toivanen et al. 2021. The second dataset revolves around the discourse that has been taking place in the legacy media sources Yle, Iltalehti and STT. This dataset consists of articles related to alcohol policy. The goal of the study is to reveal, which features perform best for classification, and does their performance differ across subtasks.

As classification features, contextual sequence representations are extracted from the finBERT language model. Topic distributions are extracted from topic models that are trained on the data. Additionally, a structural featureset developed in Toivanen et al. 2021 is utilized. These structural features consist of different markup features of the articles, such as distances between tags and image sizes. The hypothesis that BERT-based embeddings could be improved upon by augmenting them with additional information is reinforced by recent good results in natural language benchmarks and tasks (Peinelt et al 2020, Glazkova 2021). By combining contextual embeddings with topics, only marginal performance increase is achieved and only in certain environments. In most instances, the combination was detrimental to performance due to increased noise in the classification feature. Nevertheless, various combinations of BERT-based embeddings, topics and structural features were found to outperform purely BERT-based classification in many subtasks. Additionally, potential future developments to achieve better classification performance are outlined.

Based on the experiments, automated frame analysis with neural classifiers is possible, but the accuracy is not yet sufficient for inferences of high certainty.

Contents

Chapter 1

Introduction

Advances in the accessibility of technology have motivated a growing interest in combining qualitative and quantitative methodologies in media studies and modern machine learning methods make it possible to gain insights from large datasets that would be impractical to analyse with more traditional methods. This thesis presents experiments in training supervised document classifiers that combine BERT-based contextual embeddings with topic information and structural feature information with a goal of improving classification performance over a purely BERT-based classifier. The hypothesis that features derived from document structure and a topic model could complement features derived through embeddings is relevant, because contextual embeddings contain mainly linguistic and distributional information up to the level of a couple sentences, whereas topic and structural information allows us to introduce document level structural information. This should be a complementary addition and useful for document-level classification tasks. The approach is tested through two test scenarios involving the classifying of Finnish countermedia articles and Finnish legacy media articles, where deep learning methods are proposed as solutions.

The countermedia dataset represents a real classification need, considering there has been an influx of terms like "fake news" and "alternative media", that have recently been used to juxtapose countermedia with legacy news sources in the hybrid-media environment. There is a growing body of evidence to suggest that the implied boundary between truthful and untruthful news may not be as clear as it would seem. Countermedia outlets, instead of publishing strictly made up content, might instead introduce biases more by means of curation of information (Ylä-Anttila et al., 2019). Countermedia publications have emerged to counter a perceived bias of mainstream news outlets (Reunanen, 2018). The countermedia plays a part in the rise of right-wing populist politics in Finland and thus remains an important focus of study (Ylä-Anttila et al., 2019). On the other hand, the alcohol political discussion in Finnish legacy media is of interest to media scholars because of the substantial public health impact of alcohol usage and the discussion around personal freedom and state control. By retrieving relevant aspects of

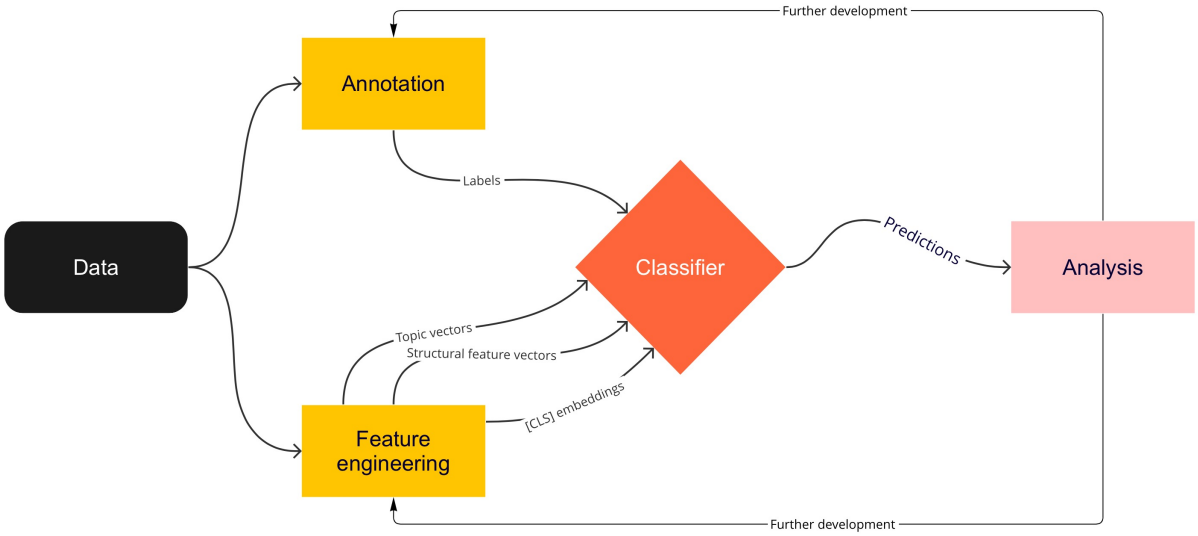


Figure 1.1: Workflow for supervised classification in a media research context

alcohol-related discussions, researchers can make more accurate quantitative inferences from large datasets.

The goal of the study is to assess the capacity of the classifier-feature combinations to capture the analytical concepts required for the study of the content objectives related to the two research projects. Both tasks are modeled as a supervised classification problem. The first task is automated frame classification which is carried out on a countermedia dataset and the second task is content detection from legacy news media sources.

Supervised neural networks learn features from data by comparing the current state of the classification to the annotated gold standard and propagating the required corrections across the network to update the weights and biases of the neural network. The objective of a supervised neural network is to learn features that enable the neural network to emulate the choices of the annotator. The supervised framework enables the researcher to evaluate the research process in terms of established theoretical frameworks that are relevant to the field of study, in contrast to unsupervised methods where the choice of algorithms determines the results to a larger degree. Supervised classification presents a good platform for combining expert domain knowledge and close reading with broader quantitative analysis. The approach is first evaluated through a test scenario involving the classifying of Finnish countermedia articles. Then a comparison of the classifiers is carried out by comparing the classification performance of the selected features between the two datasets and tasks. The thesis aims to answer the following research questions:

* RQ1 Can document classification that utilizes BERTs [CLS] embeddings be improved upon by augmenting the [CLS] embeddings with topic information and structural features?

* RQ2 Which classification features work best for the two tasks?

Answering these questions entails acquiring document-level classification features from a pre-trained language model, a topic model and raw html files. In the results I will show how the different classification features and their combinations perform in the two tasks. Based on the results, I will also reflect on the limitations of the methodology and briefly discuss the possible future avenues and possibilities for document classification in a computational social science environment.

Chapter 2

Data

2.1 Countermedia dataset

The first dataset is a corpus of countermedia articles gathered and annotated by Toivanen et al. (2021). It contains 37 185 articles from the Finnish countermedia publication MV-lehti, gathered from their website between 8/2014 and 3/2018. The dataset was chosen to develop the classification performance of automatic classifiers in a task where the classification should be based on abstract definitions of "frames".

MV-lehti is the most popular and by volume the largest countermedia publication in Finland: in March 2018 the website had over 800 000 visitors and was estimated to reach 5% of the population weekly (Tuomola, 2018; Reunanen, 2018). Its content illustrates an anti-establishment, anti-immigration, right-wing agenda. The period the data represents is characterized by an active publishing schedule on matters related to immigration policy and the perceived societal issues around immigration and refugees. MV-lehti appeared in the public sphere following the rise of an anti-immigration movement that was triggered by the so-called “european immigration crisis” that started in 2014-2015. The publication aims to polarize the attitudes of like minded individuals towards specific topics, such as immigration and refugees and it is usually categorized as part of a group of Finnish countermedia outlets that have a strong anti-immigration stance (Tuomola, 2018). These publications curate information from the news flow and modulate it to fit their agenda (Reunanen, 2018). The data contains the titles, main bodies of text contents including html markup, links to the articles, release and modification times, sources used and author id.

Annotation

To facilitate an analysis of the curation of information in MV-lehti, a subset of 997 annotated articles were chosen for training and evaluation. These articles have been annotated into three

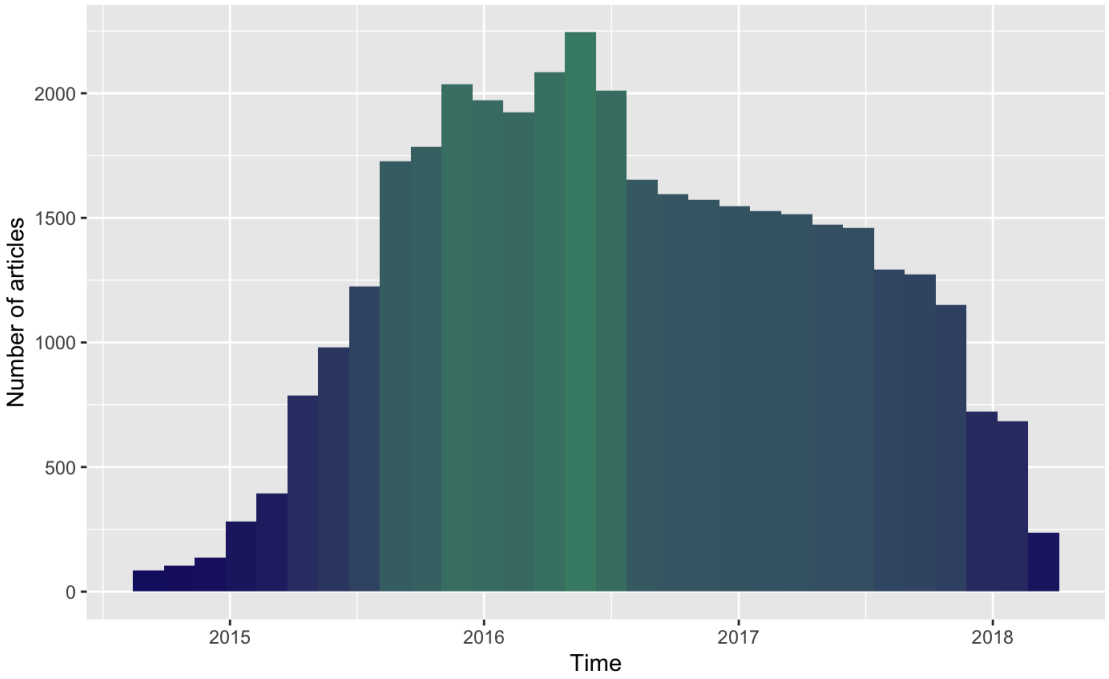


Figure 2.1: The distribution of the countermedia articles in time. The peak coincides with the height of reporting on the refugee crisis and wanes as mainstream media shifts its focus from the refugee crisis to other subjects, and diminished further in 2018 when editor Ilja Janitskin was faced with charges of ethnic agitation and defamation.

classes by Toivanen et al. (2021). Each article contains a reference or a hyperlink to mainstream media sources and the three classes represent three different ways to re-contextualize information from external sources. The goal of the countermedia-classifier is to learn the distinctive features of the classes and classify unseen articles reliably based on the learned features.

The first category is criticisms of mainstream media. Articles in this category contain criticism of mainstream media, whether it is accusations of lack of coverage on certain subjects or slandering a particular journalist. The annotation guidelines were such that if there was any kind of mainstream media criticism, it would override other features present in the article. The second category contains copies from mainstream media. The articles are mostly recycled content, but some contain comments and novel use of typographical effects such as bolding, exclamations and quotes. The final category consists of texts that use external sources to reinforce a narrative that is contrasting in its nature compared to the source article content. For example, a news report about adding resources to handle an increased amount of paperwork related to immigration is used in an article that claims all the money is going to the "pockets of lawyers".

The copies class and the original content class have distinct prototypes: most of the articles categorized as copies are more representative of traditional news text, while the original content articles are longer, free-form and contain a lot of html markup, creative language use, images, and links to external sources. The class distribution of the dataset is unbalanced and most of the

content of the dataset consists of copies from mainstream media:

```
class 1: criticism of mainstream media, 81 articles
class 2: copies from mainstream media, 770 articles
class 3: original content, 146 articles
```

The texts in the dataset vary in style: there are different lengths of articles from reposts of videos with only a few sentences to long analyses and rants. Article types range from police reports through copies from other media outlets to blog-like texts and fictional pieces. Editorial commentaries are frequently mixed within copies of texts from other news sources. The publication is out of reach of the institutional regulation that has an influence on traditional media outlets. Like many similar countermedia outlets, MV-lehti does not conform to the ethical guidelines and responsibility principles for journalists that are enforced by the Finnish Council for Mass Media.

2.2 Alcohol policy dataset

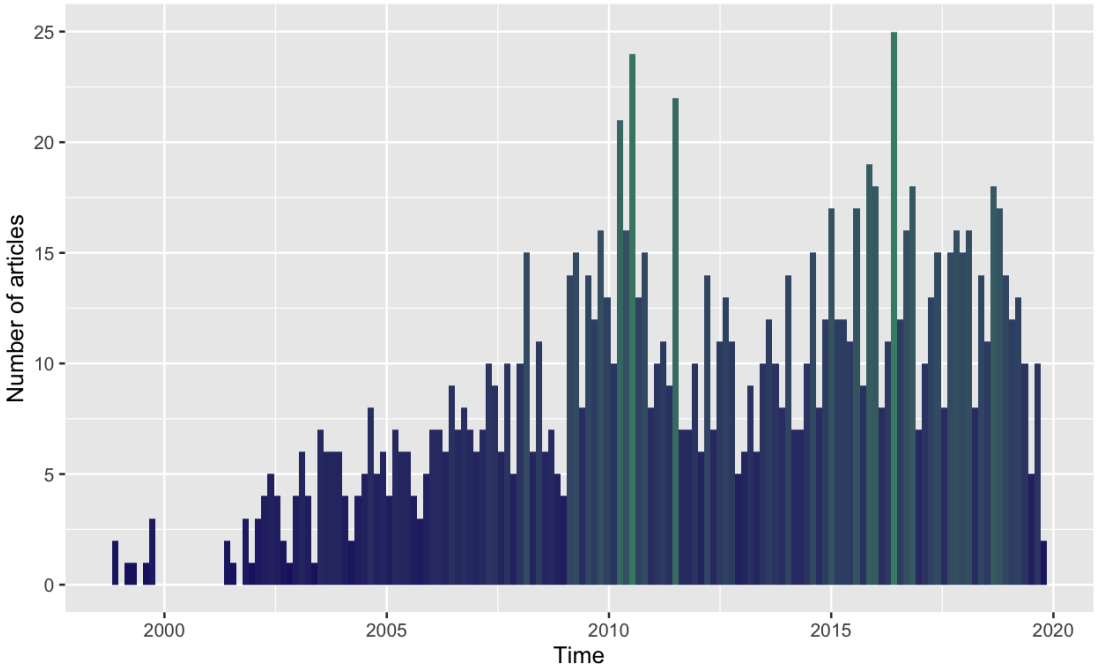


Figure 2.2: Distribution of alcohol policy articles in time. The first peak coincides with the proposals of the Finnish parliament to raise taxation of alcohol, soft drinks and candy. The second peak occurs when the increases are put to practice

The alcohol policy dataset is an annotated sample of 1600 articles from a 33 902 -article legacy media dataset, where each article mentions the Finnish word for alcohol. The articles

were gathered from YLE, STT and Iltalehti as part of an ongoing Flows of Power investigation into the discourses on alcohol policy in legacy news media platforms. From the alcohol-related articles, the media researchers of the project found three subclasses of interest:

1. Alcohol legislation
2. Markets and research on alcohol consumption
3. Prevention and treatment of alcohol harms

At the start of the project, 1500 articles were annotated by three annotators and a 100 articles were annotated by 10 annotators. All the annotators were recruited from the University of Helsinki. At the time of writing, a proper 3-class classification experiment was not possible to implement, as the inter-annotator agreement was too low to reasonably attempt computational classification and further development of the annotation scheme was necessary. A subset of 1124 articles that had perfect mutual agreement on a binary alcohol policy/not alcohol policy basis was extracted. This gave me reasonable confidence to use the data to train the classifier and compare the classification features across two datasets. The more fine-grained classification could be examined in the future.

Chapter 3

BERT

The language model used to extract contextual sequence representations from the articles is BERT, which is a state-of-the-art language representation model built using a novel pre-training framework. Google introduced BERT as part of their search engine in 2019 to improve its ability to understand more intricate queries. The pre-2019 search engine relied on matching keywords, which led to having to learn a specific "keyword language" to get good results, but BERT is capable of extracting contextual knowledge and subtle nuances from more "natural" search queries.

3.1 Background

BERTs architecture is based on a parallelizable encoder-only transformer architecture that relies on self-attention mechanisms to calculate different kinds of dependencies between the input and output (Devlin et al., 2019). Through its pretraining objectives, BERT produces deep, nondirectional language representations called contextual embeddings. These embeddings are token-level language representations that encode contextual information by learning dependencies through a stack of multi-headed self-attention layers. The resulting language representations are commonly utilized in a semi- or fully supervised classification task. This procedure is called transfer learning: information that has been accumulated from a particular learning process is transferred for use in other tasks that require computational language understanding.

BERT broke the tradition of pretraining sequential and window-based language models like recurrent neural networks and convolutional neural networks using next token prediction, and introduced masked token prediction as a pretraining objective. This removed the need for sequential processing. The biggest bottleneck in recurrent language models is their sequential nature, where computation is factored along positions of sequences and aligned with steps in computation time. Sequential models generate a sequence of hidden states as a function of the previous state and the current input. This leads to memory issues with longer sequences and

makes the computations difficult to parallelize (Vaswani et al., 2017).

Also, contrary to convolutional neural networks such as ConvS2s or ByteNet, where the number of calculations required to relate input and output increases linearly or logarithmically with distance, each BERT input has the same distance to every output (Vaswani et al., 2017; Devlin et al., 2019). BERTs architecture is faster for calculating dependencies between inputs and outputs and this is because of the way attention is implemented in the system Devlin et al. (2019).

3.2 Multi-headed self-attention

Attention is a way to map dependencies between different aspects of the inputs using queries and key-value pairs and it has been previously used in sequential architectures, but BERT is the first model where it is implemented to transformers in a multi-headed, self-attending configuration (Vaswani et al., 2017; Devlin et al., 2019). BERT consists of 12 self-attention layers that contain 12 parallel self-attention functions, or "heads" that in turn consist of a layer normalization, a self-attention module, and a feedforward connection (Devlin et al., 2019). This parallelized self-attention is called multi-headed self-attention. In the self-attention layers, the key, value and query vectors are projected with learned linear projections into 64-dimensional subspaces that can encode different aspects of the input (Vaswani et al., 2017; Devlin et al., 2019). Since the 12 self-attention layers are parallel and independently parametrized, the model is able to attend to many representational subspaces of the input sequence at different positions (Vaswani et al., 2017). In practice this means that given a sentence, multi-headed attention enables the model to simultaneously learn many different kinds of dependencies between tokens. It has been hypothesized that different attention heads may encode linguistically interpretable information, but the subject is still contested (Rogers et al., 2020; Jain and Wallace, 2019). Finally the 64-dimensional outputs of the self-attention layers are concatenated into a 768-component contextual embedding.

3.3 Pretraining

Devlin et al. (2019) also introduced a new way to pre-train self-attention based transformer models, and BERTs architecture enables the fine-tuning of the parameters of the model by modifying the final output layer, called the classification head. The model is pre-trained with two pre-training tasks: masked word prediction and next sentence prediction. In masked word prediction, the training objective is to predict words that are masked or obfuscated from the input. This is in contrast with how sequential models are pre-trained, where the objective is to predict the next word in the sequence. The masked language objective makes it possible to learn

the full context at once. In the next sentence prediction task, the model is given two sentences A and B, where 50% of the time B is a random sentence from the dataset and 50% of the time a sentence that follows A in the dataset. Next sentence prediction was found to further improve performance in question anwsering and natural language inference benchmarks (Devlin et al., 2019).

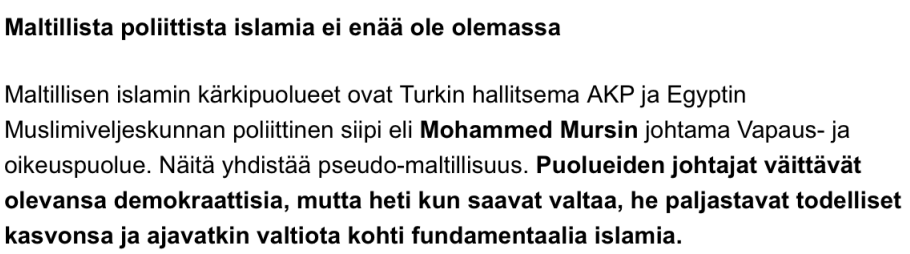
Chapter 4

Feature engineering

The features extracted from the data for the classifiers are the [CLS]-embeddings from BERT and topic vectors that are obtained by training a topic model on the data. Additionally, for the countermedia case, a structural featureset engineered by Toivanen et al. (2021) was used.

4.1 Structural features

The use of structural features is motivated by the diversity of stylistic and structural features present in the MV-lehti articles.


Maltillista poliittista islamia ei enää ole olemassa

Maltillisen islamin kärkipuolueet ovat Turkin hallitsema AKP ja Egyptin Muslimiveljeskunnan poliittinen siipi eli **Mohammed Mursin** johtama Vapaus- ja oikeuspuolue. Näitä yhdistää pseudo-maltillisuus. **Puolueiden johtajat väittävät olevansa demokraattisia, mutta heti kun saavat valtaa, he paljastavat todelliset kasvonsa ja ajavatkin valtiota kohti fundamentaalia islamia.**

Figure 4.1: An example of the variation in the use of html markup: simultaneous use of bolding as the header, to mark a named entity and as rhetorical emphasis.

Because finBERT is not pretrained on html data, it cannot parse the raw html articles in a way that could make it possible to utilize these features. To model the varied structural content of the countermedia articles, Toivanen et al. (2021) developed a structural featureset that is utilized in this thesis. From the data, Toivanen et al. (2021) named 80 structural features that represent features such as distances between various tags, amounts of tags, links, sizes of pictures, length of an article, use of special characters, such as "!?", "!!!" and "...". It was shown that structural features are effective in classifying articles from the countermedia dataset using a random forest classifier (Toivanen et al., 2021).

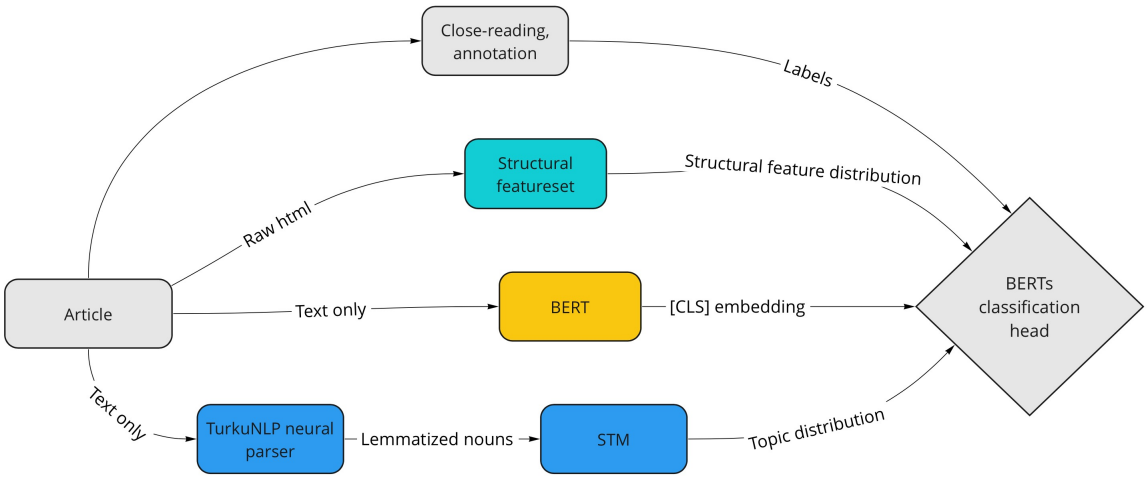


Figure 4.2: The feature extraction process

4.2 [CLS]-embedding

The [CLS]-embedding represents the [CLS]-token which is a special token prepended by BERT to all input sequences to facilitate pre-training on next sentence prediction. This special token is not masked during pre-training, so through self-attention it is updated and learned in all training instances. The [CLS]-embedding thus becomes an aggregated representation of the input sequence and it has been used for text classification tasks such as toxic content detection or detection of extreme sentiments (Xiang et al., 2021; Jamil et al., 2022). The [CLS]-embedding for a given input sequence is extracted from the last hidden layer of a [CLS]-token. What kind of information exactly is encoded into the [CLS]-embeddings is not yet fully understood, but the embeddings are commonly used as an aggregated representation of the input sequence (Rogers et al., 2020).

4.3 Topic vectors

Topic models are a type of generative latent variable model, where the latent variables are multinomial distributions of word co-occurrences that are referred to as "topics". Topic models consist of a user-specified number of topics as a distribution over the vocabulary, and a distribution over the topics for all documents (Chang et al., 2009). A generative process is described for each document, where each document is modeled as a multinomial distribution over the topics, which in turn are a multinomial distribution over the vocabulary (Roberts et al., 2019). The structural topic model has been developed for the use of social scientific research and has been useful for various tasks such as analysis of high court judges' tweets (Curry and Fix, 2019)

to comparative politics (Lucas et al., 2015). In studies that utilize topic models, it is typically assumed that the latent space is semantically meaningful, but in this work the models are not used to analyze the structure and semantics of the dataset, but rather to investigate whether topics could act as a proxy for the frames described in Toivanen et al. (2021) and for the alcohol policy task. The topic model used for the countermedia task is a structural topic model, trained using the stm package for R and for the alcohol policy task, an LDA model was used.

4.3.1 Countermedia task

The dataset was first processed through the TurkuNLP neural parsing pipeline. The TurkuNLP Group is a group of researchers based in University of Turku and they are the leading developers of NLP tools for the Finnish language. Their pipeline includes tools for syntactic and semantic analysis of Finnish such as tokenization, sentence splitting, lemmatization, morphological tagging and dependency parsing. The pipeline represents the state of the art in Finnish lemmatization and is as of writing ranked 1st on lemmatization on the ConLL-18 shared task on parsing universal dependencies (Kanerva et al., 2018). The parsing process ran on CSC:s Puhti supercomputers small 1-node partition as a batch job which was coordinated using Slurm. 200G of memory was allocated for the CPU and with these parameters the processing was complete in 8 hours.

From the processed data, lemmatized nouns were extracted to be used as features for building the topic model. Nouns are typically used to name objects, entities and ideas and would thus be a good feature to use for modelling topical content in texts. In this case, the decision to use only lemmatized nouns was made based on close reading of the dataset and looking at word frequency counts, where I found that there is a large amount of weapon and war vocabulary, which reinforced my thoughts behind using nouns as a proxy for the different frames. Lemmatized nouns were chosen to have a purely lexical classification feature, to contrast with BERTs contextual embeddings and the structural feature vectors that take into account a degree of syntactic dependencies and html markup. Including verbs, inflections and function words in the topic model would have by proxy introduced more syntactic information into the signal. The topic model for the countermedia task was trained on the full 37 185 article dataset.

The lemmatized articles were projected into a “tidy” format with the R tidyverse library and the two topic models were trained using the stm package for R. One model was trained with K = 200 and another one with K = 76, which was determined by the Lee & Mimno algorithm, which optimizes K for topic specificity and topic dissimilarity (Mimno and Lee, 2014). Good topics should balance both: high specificity results in specific topics, but in an extreme case an article would be composed of only one topic and there would be as many topics as there are articles. Low dissimilarity on the other hand means a low number of topics, which usually leads to high frequency words being top contributors to most topics (Mimno and

Lee, 2014). The models were initialized with spectral initialization, which is a way to reduce the number of computations to find reliable anchor words for calculating the topics and has been found to outperform previous initialization methods (Roberts et al., 2019). Two topic models were trained, one with more fine-grained topics and one with broader topics, as there is no established method to arrive at a value of K that will produce the best outcome for a given task. The broader, optimized topics did not improve classification performance, and for clarity the 76-topic classifier is left out of the final evaluation. No metadata covariables were utilized in modelling to facilitate a fairer comparison with the LDA model used for the alcohol policy task as LDA is incapable of utilizing metadata covariables.

4.3.2 Alcohol policy task

For the alcohol policy task, I used an LDA model that had been trained on 145 289 articles of Helsingin Sanomat articles for earlier research. The model was chosen, because the Helsingin Sanomat dataset is more representative of the legacy media style and content. As a usable model was already available, training a new one was deemed unnecessary. From the topic models, document-wise topic distribution vectors were extracted to be used as classification features. Like the countermedia topic model, the alcohol policy topic model was also built from lemmatized nouns and the data had been processed by the same TurkuNLP pipeline.

Chapter 5

Experimental setup

5.1 Hardware and hyperparameters

The experiments were carried out on CSC:s Puhti supercomputer, on a single node. Each node ran two Intel Xeon "Cascade Lake" processors that have 20 cores each running at 2.1GHz. I used a gpu partition that uses a single Nvidia Volta V100 GPU with 32 GBs of memory. The code for the experiments was written in Python and interfaced with Jupyter Notebooks with data-wrangling and visualisations done in R. The code is available for review in Github. The uncased, 12-layer, 768 hidden, 110M parameter PyTorch HuggingFace implementation of the finBERT-base model was used to train each classifier. The models were fine-tuned over 400 epochs with a batch size of 32 and a dropout probability of 0.1. BERTs weights were frozen as recommended for feature-based approaches by the original authors and to counter severe overfitting (Devlin et al., 2019). The time to train a classifier with these parameters is around 1.5 hours.

In absence of a 3-class alcohol policy annotation, in order to compare the performance across the two tasks, the countermedia minority classes "criticism of mainstream media" and "original content" were collapsed into one so that copies from mainstream media remained an independent class.

Cross entropy and binary cross entropy loss functions are used for the 2-class and 3-class tasks respectively. Class weighing was modified to make minority class misclassifications more costly. The data was split into 80/20 training and evaluation sets using the train_test_split method from the scikit-learn library for Python.

The hyperparameters were identical for all classifier-feature combinations.

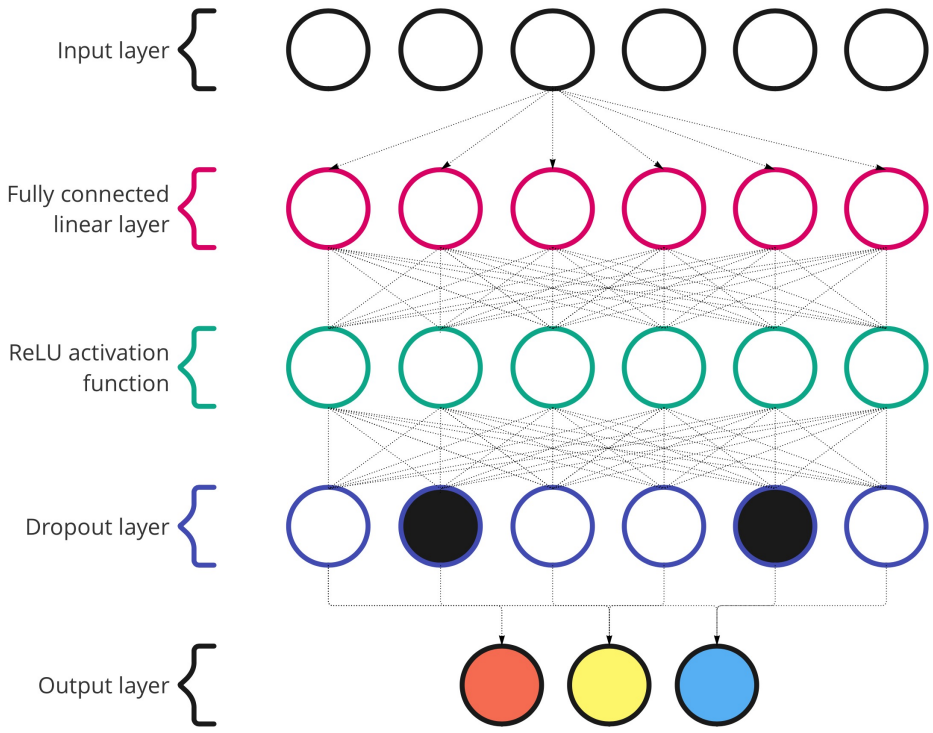


Figure 5.1: Structure of the classifier head up to the output layer, with the connections of a single input node illustrated

5.2 finBERT

For the language modelling, I used finBERT, which is a BERT model pretrained on 98M Finnish documents (Virtanen et al., 2019). The model represents the state of the art in Finnish language modeling and it is trained on texts that range from unrestricted internet crawls of the Finnish internet, through online-discussions to legacy news media articles (Virtanen et al., 2019). BERTs code was modified to accommodate texts that are longer than 512 tokens. This was done by splitting longer articles into snippets of 512 tokens and tracking the input ids, token type ids, attention masks and the corresponding topic and structural feature vectors over the subsequent snippets of each article. This made it possible to use the article-specific feature vectors to classify multiple snippets of a single article.

5.3 Classification head

The classifier itself is a multi-layer perceptron built on the BertForSequenceClassification class from the PyTorch Huggingface library. The custom MLP replaces the stock linear classification head of the class. The forward method of the class is responsible for pooling the CLS embeddings from the last hidden layer of BERT, feeding them into the classifier head and cal-

culating loss. The method was modified to use not only BERT-based encodings, but also topic and structural feature vectors and any combination thereof. Multiple configurations of hidden ReLU layers and linear transformations were experimented with, but the architecture presented above provided the best results along with a fast training time and reliable loss behaviour over all the experiments. The classifier has a fully connected linear input layer, a hidden ReLU layer, a dropout layer and an output layer. The size of the fully connected linear layer varies with the input: 768 for the [CLS] inputs, 200 for the topics and 80 for the structural feature vectors, and the sum of these in the case of any combined features. In the three classifiers where combinations of features were used, the feature vectors were concatenated before forwarding the full feature vector into the classification head. All the models that are presented were chosen based on lowest loss on the evaluation dataset.

Chapter 6

Results and analysis

In this chapter, I will present the class-specific recall, precision and microaveraged f1 scores for each classifier. Then, I will go through the notable error modes for the classes and conclude with further error analysis. Finally, the results and error analysis for the countermedia-alcohol policy comparison are presented in section 6.2.1.

6.1 Results for the countermedia multiclass classifiers

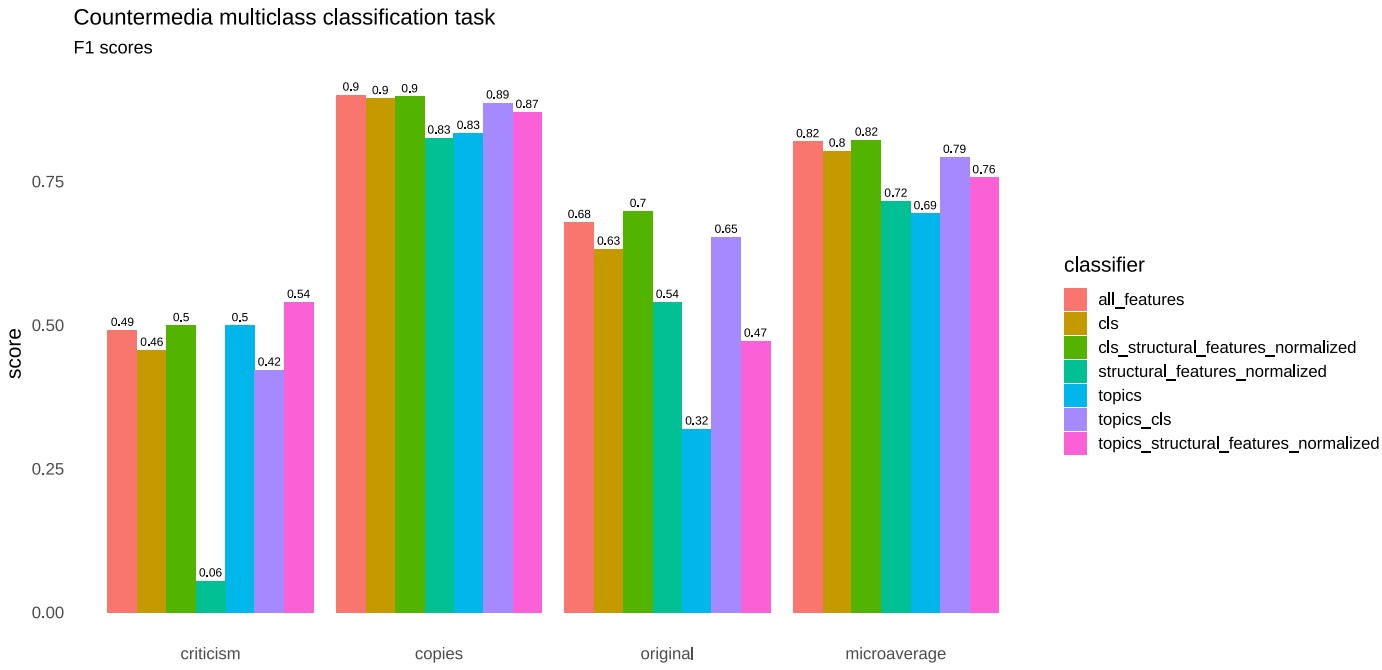
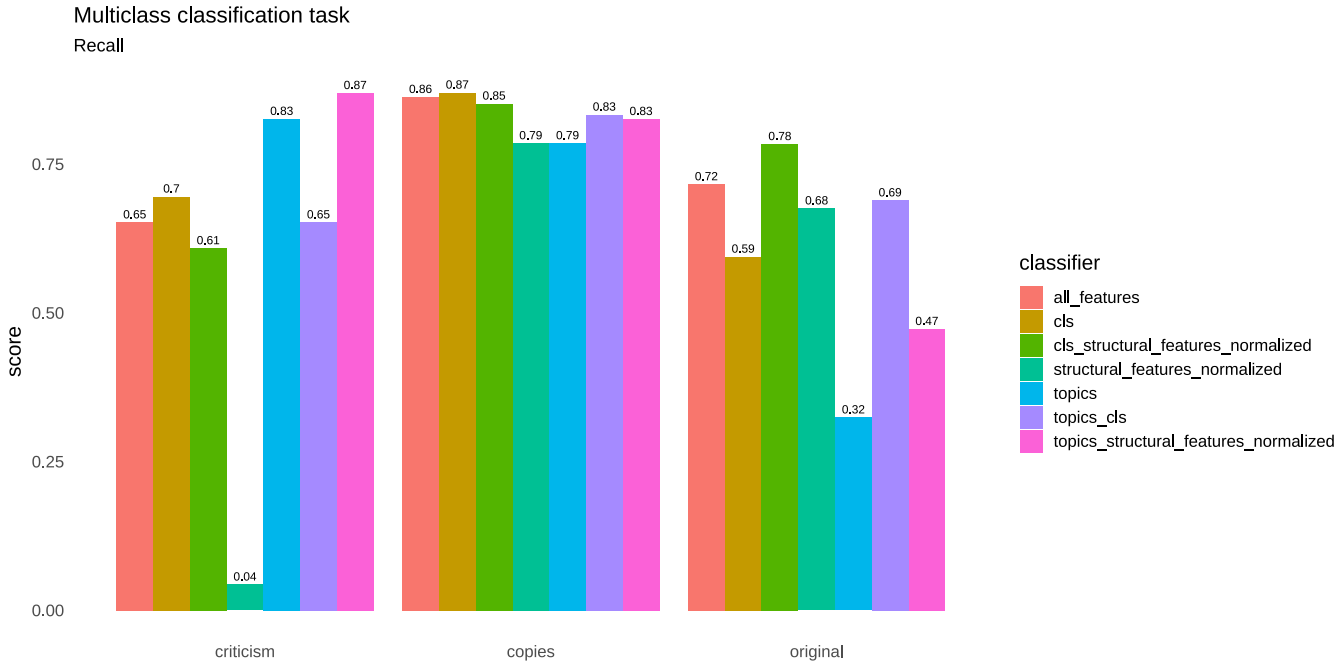
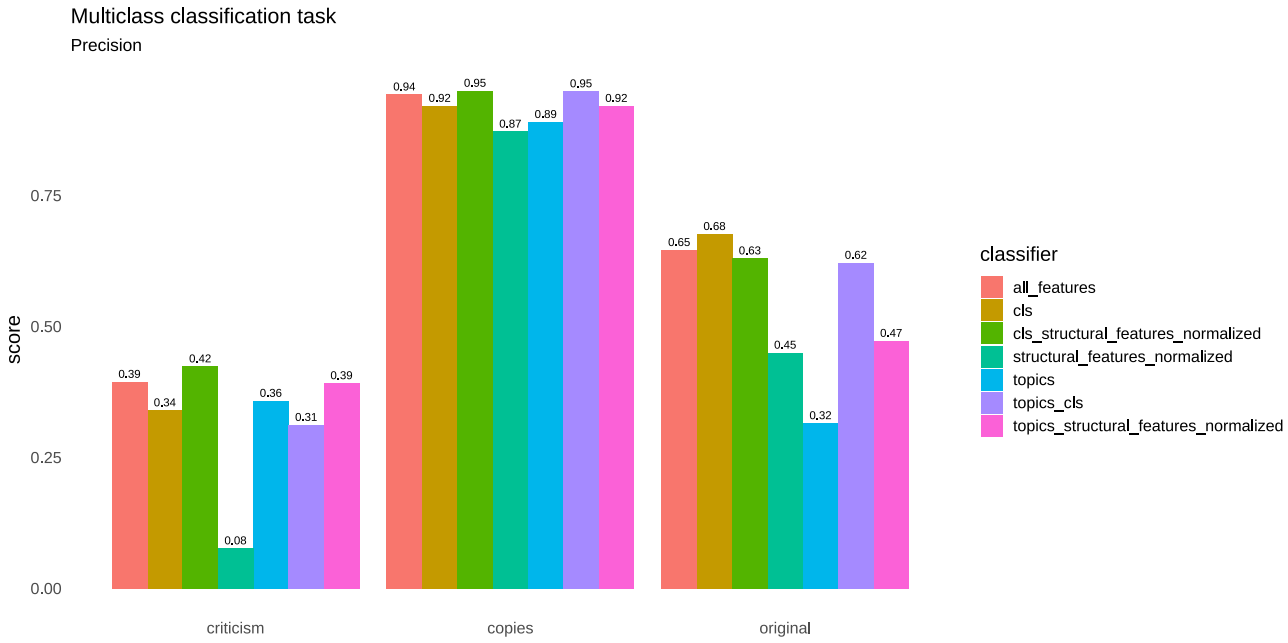


Figure 6.1: Class-wise f1 scores for the 3-class countermedia classifiers





6.1.1 Overall accuracy

The best overall classification accuracy was achieved with a combination of structural features and [CLS]-embeddings. This combination yielded an accuracy of 0.82, a 2-point increase over the [CLS]-baseline. A combination of all features failed to bring a performance increase over the [CLS]-baseline and the [CLS] + structural feature classifier , but [CLS]-embeddings on their own performed reasomably well, with an accuracy of 0.82. Topics and structural features per-

formed similarly (0.72 vs 0.69), but structural features showed significantly better performance on the original content class (0.52 vs 0.32), while topic-based classification performed better on criticisms of mainstream media (0.5 vs 0.06). Across all classes, adding structural features to [CLS]-embeddings was more beneficial than adding topics to cls embeddings.

Classification performance for the minority classes

The best classification performance for the criticism class was achieved by a combination of topic and structural information. Topics and structural features also have the highest recall in the class (0.87), but at the cost of lower precision (0.39). As a standalone feature, topics performed better than the [CLS]-embeddings. For the original content class, the best classifier utilized [CLS] embeddings and structural features. This classifier outperformed the [CLS] baseline by 7 points (0.7 vs 0.63). As a standalone feature, structural features considerably outperformed topics in the original content class (0.54 vs 0.32). Notably, topics and structural features have very high recall for the criticism class (0.87), with similar precision scores to the [CLS]-embeddings. In the original content class, purely [CLS] based classification yields the highest precision at 0.68 points. Conversely, for the original content class, the highest recall (0.78) is produced by [CLS] embeddings combined with structural features.

6.2 Analysis

Even though the accuracies of the classifiers are relatively similar, there are differences in the way the classifiers work. The sets of misclassified articles were largely different for the classifiers, although there were subsets of articles that were misclassified by multiple classifiers. Also, for each class, there were 3-9 articles that were misclassified with all classifiers.

Most of the performance difference between topics and structural features in the original content class comes from articles that are in English. Structural features, working on a non-lexical level, managed to classify these correctly, while there were clearly too few English articles in the dataset for the topic model to resolve correctly. The structural features often (14 out of 22 misclassifications) mistake criticisms to be copies. Even though all the feature vectors were normalized to the same range of values from 0 to 1, the normalization was still too crude. The feature vectors were normalized separately in their own respective pipelines and concatenated before feeding into the classification head. What followed was that longer feature vectors tended to have smaller values for the components. Combining a 768-length vector with a 200-length vector thus made learning less efficient as the components with larger values would be more affected by a backpropagated correction than components with smaller values. This would partly explain why adding topics to [CLS] embeddings generally makes the performance of the classifier worse, as the larger values of topic vectors would negatively affect the loss

behaviour. This effect is also compounded by the fact that both topics and [CLS]-embeddings encode a degree of probabilities of lexical co-occurrence. When topics were combined with [CLS] embeddings, some English articles that the [CLS]-classifier classified correctly became misclassified, which would imply that there are instances where topic components override or introduce noise to the signal.

Meanwhile, structural features encode non-lexical characteristics of an article and thus combining them with lexical features enhances the performance of the classifier. In future, it should be examined if the performance decrease is indeed due to issues with normalization, by developing a more sophisticated normalization system, or is it related to the hypothesis that both features encode similar co-occurrence patterns in a way that introduces noise to the signal.

Lemmatized nouns are also probably not the best features to train a topic model for this task, considering that the countermedia dataset is topically centered in matters related to the ills of immigration. Stripping all other linguistic aspects likely levels the articles to an extent where it is harder to separate more nuanced classes such as the frames that are the target of this task. Consider this translated example sentence of a criticism of mainstream media from the countermedia dataset:

> "Yle, Helsingin Sanomat and the Finnish christian media outlets have not reported on the incident which was covered by the German quality publication Welt and MV-lehti."

The media criticism in this sentence is not obvious from lemmatized nouns, but rather encoded in a more complex construction "X has not reported on Y, but Z has". The different types of recontextualization may well be encoded in aspects such as dependency structures, use of verbs or modifiers. The good performance of BERT might in part be explained by the fact that it does seem to encode some degree of these kinds of more complex dependencies (Rogers et al., 2020).

The results for the criticism class correlate with the amount of training examples available. The class had the least amount of training data and the performance of the classifiers was lowest. The structural feature classifier almost completely fails to predict the criticism class. Most of the misclassifications made by the structural feature classifier were predictions towards the copy class. The averaged cosine similarities between true copies, true criticisms, and misclassifications toward the copy class showed that the structural feature distributions of true copies and misclassifications towards copies is very similar. This behaviour shows that most articles in the criticism class share structural features with the copies class. This makes sense, as most of the criticisms are copies from other media but with some added critical comments.

What was defined in the annotation guidelines as a “critique of mainstream media” was an overriding feature during the annotation process. The guidelines stated that if an article

contains critique of mainstream media, it is to be classified as a critique of mainstream media regardless of other features present. This leads to a situation where one class contains subsets of articles that might reasonably belong to other classes; the level of abstraction is different for the criticism class. Critique articles were frequently misclassified to be copies due to them being mostly similar to them. A single sentence of direct critique in an article was not enough for a correct automatic classification, even though it had been clear for a human interpreter. This introduces complexity, and along with a very low amount of data available to learn, the classification performance remains poor.

Copies from mainstream media are also shorter in length and maximum structural feature use is lower, while original content articles are longer and richer with structural features. There is a subset of critique articles that utilizes structural features in a way that is characteristic of the copies class. The misclassifications towards the copy class were found to be remarkably similar to true copies and very different from true criticisms. Most of the criticism misclassifications were criticisms that were predicted to be copies.

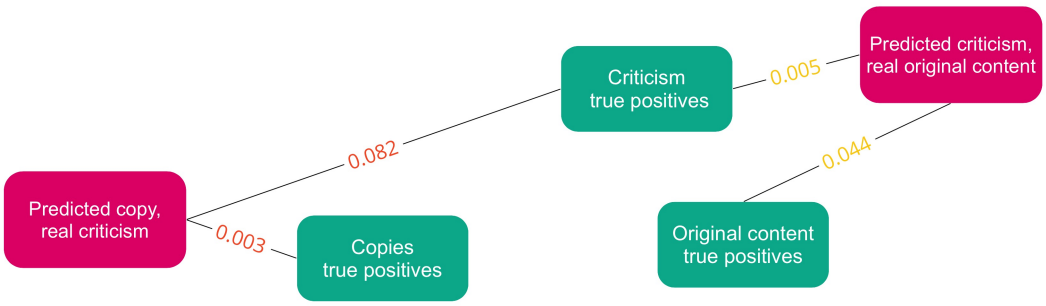


Figure 6.2: Averaged cosine distances of structural feature distributions for two error classes with the largest difference of distance between the true and predicted classes

Another problem is overlap in the structural features between classes. Many copies from mainstream media share structural features with the minority classes in such a way that the structural feature-based classification misclassified the copies as original content or criticism. A typical original content article exhibits a special structural feature distribution, with many pictures and more variance in different markup features, but if a copy article looks too much like original content, it gets misclassified.

The overlap, where two classes share similarities is also compounded with unbalanced class distributions that are known to introduce bias towards the most represented class (Santos et al., 2022). In this case, it might well be that class overlap presents a bigger problem than the class imbalance, as the feature distributions of misclassified articles are more similar to the feature distributions of the predicted class than the gold standard.

Two error sources are related to articles that exceed the input limitation of BERT. An article

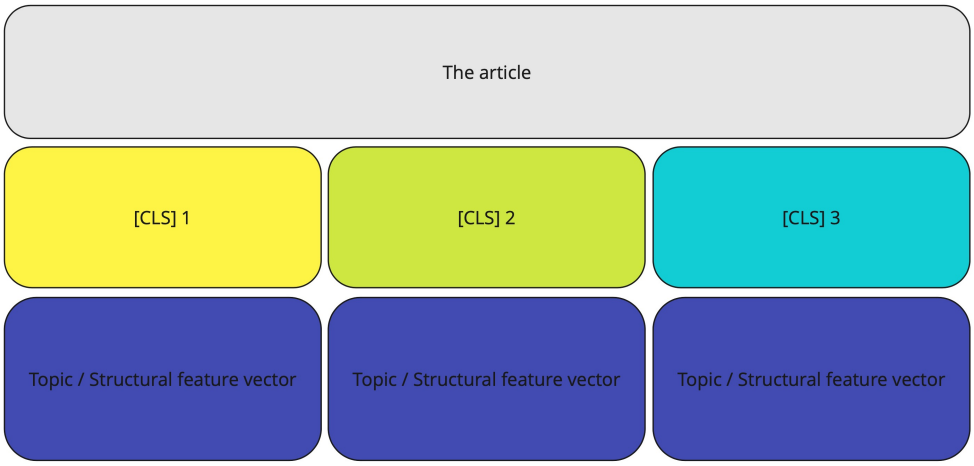


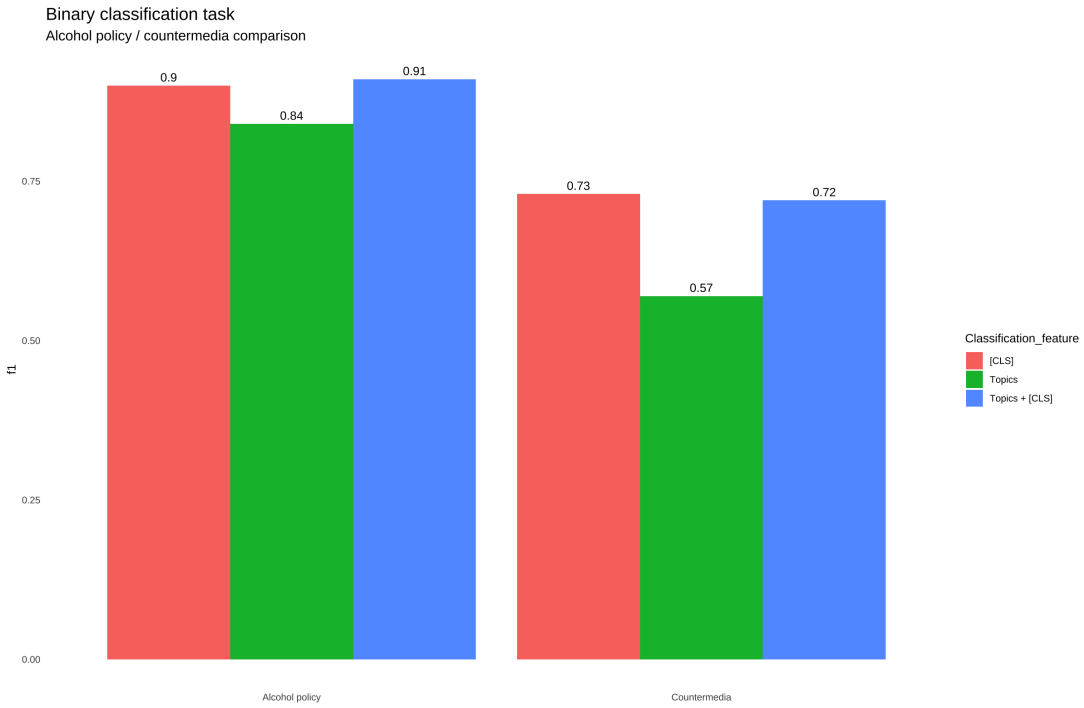
Figure 6.3: The structural feature and topic vectors represent the entire article, while [CLS]-embeddings represent separate snippets of 512 tokens.

that exceeds the 512 token limit will often be labeled differently to the main body of the article, because BERT considers the snippets of articles to be independent, whereas the non-BERT feature vectors are representations of the entire document. Consider an article of 1500 tokens. It is split into 3 snippets that each have a unique [CLS] embedding that in many cases yields multiple separate classifications. One snippet of the article might be classified correctly, while the rest might get misclassified. The topic and structural feature vectors on the other hand are extracted from the full 1500 token article, and each 512 token snippet carries this full representation on which the classification is carried out. As a result, in the case of longer articles, if the representation is wrong for one snippet, it is wrong for all snippets and conversely, if the classification is correct for one snippet, it is correct for all. It might be possible to solve this in future using some sort of a voting function, for example based on the aggregate of the logits for the snippets of the article. The TurkuNLP Group is also working on a Longformer language model that could process sequences longer than 512 tokens.

Finally, the classifiers seem to hit a performance ceiling at around 0.80, which coincides with the available inter-annotator metrics. Although there are error sources unrelated to annotation, it is unlikely that the classifiers could exceed the accuracy of the annotator.

6.2.1 Countermedia-alcohol policy comparison

The classifiers perform better on the alcohol policy task. The reasons for this are related to the data, annotation processes and the task itself. There are large differences between the datasets. Legacy media outlets in Finland are regulated on multiple levels, from markets to laws and they are also bound by a commitment to the guidelines of journalists. The countermedia on the other hand resigns from these guidelines and position themselves as opposite to established media.



As a consequence, the articles in the countermedia dataset exhibit a much wider range of self-expression. It is possible that the use of language in MV-lehti has diverged from norms to a point where the pretraining of BERT does not catch all the nuances. As Ylä-Anttila et al 2019 note, MV-lehti and other countermedia employ a particular kind of discursive political style that is very different from that of conventional media. For example, there are novel uses of the nouns “moniosaaja” (fin. a versatile person who excels at many tasks) and “nuorukainen” (fin. a youngster). In MV-lehti, “moniosaaja” and "nuorukainen" are not used to refer to their usual referent, but to a demographic of dark-skinned refugees and immigrants who have been tried for or accused of sexual harassment, rape or assault. Although finBERT is trained on the Suomi24 corpus that does include similar use cases, it remains unclear whether there are enough examples for pretraining to capture the meaning of such alternative uses. To understand the references requires at least some degree of cultural knowledge related to MV-lehti or comparable sources.

Colloquial language and typing errors are also more common, as are variations in the use of typographical effects such as bolding and italicization. In legacy media, special typographical emphases are tied to specific functions, for example bolded text is usually reserved to highlight named entities such as names of people and organizations, but in the countermedia dataset, it is also commonly used for creative emphasis and expression.

The inter-annotator scores were significantly higher in the alcohol policy task. The alcohol policy dataset was annotated by three people and only articles that had perfect mutual agreement were picked for training and evaluation. The countermedia dataset on the other hand was annotated by one person who did an inter-annotator test with themselves using the same anno-

tation guidelines. This resulted in a "self-agreement" of 0.8, so out of 10 original annotations, 2 were annotated differently the second time around. Most of the misclassifications were related to ambiguities regarding what is considered criticism of mainstream media or a copy. This ambiguity in the original definitions of the categories might propagate through the neural classifier into the classification results and introduce a ceiling for maximum classification performance. Classification is a hard problem even for humans, and neural networks will not automatically fix problems that are present in the data (Chang et al., 2009; Nelimarkka).

The annotation frames for the two tasks are also different and based on different analyses. In the alcohol policy case, the task to differentiate alcohol-policy related material from non-related material is topical and the level of abstraction is lower, compared to the countermedia task where it is not topics that need to be differentiated by the classifier, but whether an article is a copy from external media or not. Topics can be quite reliably extracted from texts using lexical signals, but recontextualization is more difficult to pin down. This echoes Ylä-Anttila et al. (2022) where they found that when frames are defined to be topical, topic models can present a good proxy for frames.

Chapter 7

Conclusions

To answer RQ1, the combination of topic information with CLS embeddings did not turn out to be beneficial for the countermedia classification task. On the other hand, the combination brought a marginal performance increase in the alcohol policy dataset, although in practice the performance increase is negligible. The structural featureset that was previously shown to perform well with a random-forest classifier performed well, and structural features are now confirmed to be helpful also to deep neural classifiers.

Although Peinelt et al. (2020) and Glazkova (2021) managed to get an increase in classification performance by adding topic information into CLS embeddings, the same effect was not achieved in these two tasks. This suggests that more detailed research is needed. Performance in synthetic benchmark tests does not always translate into more complex real world scenarios, and machine learning technologies such as the one used in this thesis are very sensitive to initial conditions that are modulated by the task, annotations and choices of hyperparameters.

Structural features were shown to synergise well with topic information and [CLS] information in multiple scenarios. Classifiers that utilized structural features outperformed the others when classifying the criticism class. This shows that there is potential in creating novel, dataset and task specific featuresets to enhance classification performance on minority classes in situations with limited and unbalanced training data. Considering the relative simplicity of the process of extracting structural features from text, compared to the computationally intensive processes required to produce [CLS] embeddings and topic vectors, this is an interesting result and suggests that there is potential in researching the use of non-linguistic features as as a proxy for complex frames of recontextualization.

A technological contribution of this thesis includes modifications made to the HuggingFace BertForSequenceClassification class, which was modified to be able to utilize classification features that are external to those extracted from BERT itself and combine them with features extracted from BERT or exclude BERT-based features completely. This provides a good experimental platform to conduct comparative studies that aim to push the state of the art even

further.

Classification on the two datasets confirmed that the [CLS]-embedding still represents a reasonable baseline for classification because of its ease of use and good performance. The results echo the fact that massive pre-trained language models produce contextual embeddings that are useful for many different kinds of tasks. On the other hand, features that required much less computation reached relatively similar scores and in some scenarios outperformed the [CLS] classification. Considering the growing environmental impact of building ever bigger language models, the results should bring hope also to those who want to be mindful of computational costs while retaining performance.

Chapter 8

Future work

Many effects of different hyperparameter choices were left unexplored. It is possible that other features than what were used in this thesis could also yield beneficial results. Preliminary inquiries into the decisive features that BERT utilized in the alcohol policy task included punctuations (Hrin, work in progress). Björklund and Zechner (2017) found that frequencies of punctuations are can act as a working proxy for syntactic information in authorship attribution, and Menon and Choi (2011) suggest that function words are better than part-of-speech information in separating authors when the topic varies. These findings suggest that commonly omitted features can carry information that is potentially useful in different classification tasks and the decision to omit certain features should be made based on case-specific analysis. Björklund and Zechner (2017) also noted that blogs contain more pronouns than other internet article types. It might make sense to test other word classes as features to improve classification performance, as the countermedia dataset contains language use that is similar to those of blogs in the sense that they are very loosely regulated. For the kind of unstandardized text found in the countermedia dataset, it might make more sense to build custom classification features that are based on a case-specific qualitative analysis of the dataset, than to use pre-existing features, as the performance did not reach levels acceptable for research. The good performance of the structural featureset also suggests that non-lexical features should be developed further for this kind of classification.

Regarding BERT-based classification features, the [CLS]-embedding might not be the best one to use for all sequence classification tasks. There is some evidence that an average over the last hidden states of the token vectors could perform better than [CLS] embeddings. During pre-training, certain layers of BERT seem to capture information that might generalize better to downstream tasks (Rogers et al., 2020). Experiments could be conducted by extracting features from only particular layers, and building a custom aggregate feature. The problem is that researchers do not quite agree on what exactly the different parts of BERT encode. Kovaleva et al. (2019) found that BERT is massively overparametrized and there is only a marginal number of

attention heads that encode linguistically relevant information.

Although the self-attention mechanism is said to be more interpretable than previous deep neural architectures, BERT is still quite difficult to interpret and even the supposed interpretability of attention modules has been contested (Jain and Wallace, 2019). It is difficult to reliably pinpoint which parts of the sequence BERT actually attends to when doing the classifications and what a correlation between attention weights and final classification actually means. In Jain and Wallace (2019), the researchers built adversarial activation patterns to study the attention patterns in BERT and found that identical attention scores can be extracted from sequences that are very different in terms of context and content. The attention patterns thus might not be explicitly reveal what the model attends to in the way the terminology is commonly used. There seems to be little correlation between attention scores and meaningful explanations for predictions. It is also difficult to conduct ablation studies where you would want to reliably filter out certain linguistic objects, such as verbs, when you do not know if you are also removing parts of encoding for some other features. In the case of analyzing the topic classifiers, I know that it is restricted to an analysis of co-occurrence probabilities of lemmatized nouns, and I can modify the feature extraction process to model topics on a combination of features, like lemmatized nouns and verbs. This way I could make inferences about what kinds of information is encoded into these word classes by comparing the results. If adding verbs to the model increased performance, it would be explicit that this performance increase came from the addition of verbs. Similar studies cannot be reliably done using features extracted from BERT. The research into BERT does not suggest that you could access a submodule of BERT that would encode a particular linguistic structure and nothing else, so fitting BERT into the world of linguistic definitions is difficult. A given linguistic feature, such as whether a word is a subject or not, is encoded in scattered parts of the network and exact features are difficult to extract. This opacity makes the results of the network hard to evaluate. When analyzing misclassifications of the topic and structural feature classifiers, it is easier to see the correlation between the results and the misclassifications, but with the BERT-based classifiers it is challenging to pinpoint why exactly a certain article was misclassified, as the embeddings encode information in a way that is not transparent enough to fit into established linguistic frameworks. The decisions that a researcher can make to filter linguistic features from BERT are quite broad: "if I disable this layer that might to some unknown degree encode "the subjectness of a word", will the system perform better?".

In future, I would like to explore the possibilities of using more interpretable classification algorithms such as the random forest, together with simple and reproducible feature engineering methods.

Chapter 9

Bibliography

Johanna Björklund and Niklas Zechner. Syntactic methods for topic-independent authorship attribution. *Natural Language Engineering*, 23(5):789–806, September 2017. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324917000249. URL https://www.cambridge.org/core/product/identifier/S1351324917000249/type/journal_article.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. Reading Tea Leaves: How Humans Interpret Topic Models. page 9, 2009.

Todd A. Curry and Michael P. Fix. May it please the twitterverse: The use of Twitter by state high court judges. *Journal of Information Technology & Politics*, 16(4):379–393, October 2019. ISSN 1933-1681, 1933-169X. doi: 10.1080/19331681.2019.1657048. URL https://www.tandfonline.com/doi/full/10.1080/19331681.2019.1657048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. Number: arXiv:1810.04805 arXiv:1810.04805 [cs].

Anna Glazkova. Identifying Topics of Scientific Articles with BERT-Based Approaches and Topic Modeling. In Manish Gupta and Ganesh Ramakrishnan, editors, *Trends and Applications in Knowledge Discovery and Data Mining*, volume 12705, pages 98–105. Springer International Publishing, Cham, 2021. ISBN 978-3-030-75014-5 978-3-030-75015-2. doi: 10.1007/978-3-030-75015-2_10. URL https://link.springer.com/10.1007/978-3-030-75015-2_10. Series Title: Lecture Notes in Computer Science.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation, May 2019. URL http://arxiv.org/abs/1902.10186. arXiv:1902.10186 [cs].

M. Luqman Jamil, Sebastião Pais, João Cordeiro, and Gaël Dias. Detection of extreme sentiments on social networks with BERT. *Social Network Analysis and Mining*, 12(1):55,

December 2022. ISSN 1869-5450, 1869-5469. doi: 10.1007/s13278-022-00882-z. URL https://link.springer.com/10.1007/s13278-022-00882-z.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, 2018.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4364–4373, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1445. URL https://www.aclweb.org/anthology/D19-1445.

Christopher Lucas, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2):254–277, 2015. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpu019. URL https://www.cambridge.org/core/product/identifier/S1047198700011736/type/journal_article.

Rohith Menon and Yejin Choi. Domain Independent Authorship Attribution without Domain Adaptation. page 7, 2011.

David Mimno and Moontae Lee. Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1319–1328, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1138. URL http://aclweb.org/anthology/D14-1138.

Matti Nelimarkka. Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja. page 29.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.630. URL https://www.aclweb.org/anthology/2020.acl-main.630.

Esa Reunanen. Uutismedia verkossa 2018. reuters institute digital news report. suomen maaraportti. 2018.

Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. stm: R Package for Structural Topic Models. *Journal of Statistical Software*, page 41, 2019.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, December 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00349. URL https://direct.mit.edu/tacl/article/96482.

Miriam Seoane Santos, Pedro Henriques Abreu, Nathalie Japkowicz, Alberto Fernández, Carlos Soares, Szymon Wilk, and João Santos. On the joint-effect of class imbalance and overlap: a critical review. *Artificial Intelligence Review*, March 2022. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-022-10150-3. URL https://link.springer.com/10.1007/s10462-022-10150-3.

Pihla Toivanen, Matti Nelimarkka, and Katja Valaskivi. Remediation in the hybrid media environment: Understanding countermedia in context. *new media*, page 26, 2021.

Salla Tuomola. Pakolaiskeskustelu MV-lehdessä: Merkityksellistämisen mekanismit ideologisissa puhuttelutavoissa. *Media & viestintä*, 41(3), October 2018. ISSN 2342-477X. doi: 10.23983/mv.75324. URL https://journal.fi/mediaviestinta/article/view/75324.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. Number: arXiv:1706.03762 arXiv:1706.03762 [cs].

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: BERT for Finnish, December 2019. URL http://arxiv.org/abs/1912.07076. Number: arXiv:1912.07076 arXiv:1912.07076 [cs].

Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. ToxCCIn: Toxic Content Classification with Interpretability, March 2021. URL http://arxiv.org/abs/2103.01328. arXiv:2103.01328 [cs].

Tuukka Ylä-Anttila, Gwenaëlle Bauvois, and Niko Pyrhönen. Politicization of migration in the countermedia style: A computational and qualitative analysis of populist discourse. *Discourse, Context & Media*, 32:100326, December 2019. ISSN 22116958. doi: 10.1016/j.dcm.2019.100326. URL https://linkinghub.elsevier.com/retrieve/pii/S2211695819301229.

Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication*, 18(1):91–112, April 2022. ISSN 1742-7665, 1742-7673. doi: 10.1177/17427665211023984. URL http://journals.sagepub.com/doi/10.1177/17427665211023984.