Bachelor’s Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Bias Detection in Czech News

Tomáš Horych

Supervisor: Ing. Jan Drchal, Ph.D
Field of study: Open Informatics
Subfield: Artificial Intelligence and Computer Science
May 2022

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Horych Tomáš**	Personal ID number:	**484011**
Faculty / Institute:	**Faculty of Electrical Engineering**		
Department / Institute:	**Department of Cybernetics**		
Study program:	**Open Informatics**		
Specialisation:	**Artificial Intelligence and Computer Science**		

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Bias Detection in Czech News

Bachelor’s thesis title in Czech:

Metody detekce vyváženosti zpravodajských textů

Guidelines:

1. Review the state-of-the-art methods of gender and media bias detection and mitigation related to machine learning algorithms for Natural Language Processing.
2. Construct Czech datasets using machine translation from available data (most likely English).
3. Analyze the qualities of the datasets.
4. Train NLP classifiers and compare the results to the original counterparts.
5. Evaluate the models on Czech news corpora supplied by the supervisor.

Bibliography / sources:

[1] Chen, Wei-Fan, et al. "Detecting media bias in news articles using gaussian bias distributions." arXiv preprint arXiv:2010.10649 (2020).

[2] Chen, Wei-Fan, et al. "Analyzing political bias and unfairness in news articles at different levels of granularity." arXiv preprint arXiv:2010.10652 (2020).

[3] Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." arXiv preprint arXiv:1908.09635 (2019).

[4] Blodgett, Su Lin, et al. "Language (technology) is power: A critical survey of" bias" in nlp." arXiv preprint arXiv:2005.14050 (2020).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186.

Name and workplace of bachelor’s thesis supervisor:

Ing. Jan Drchal, Ph.D. Artificial Intelligence Center FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **27.01.2022** Deadline for bachelor thesis submission: **20.05.2022**

Assignment valid until: **30.09.2023**

Ing. Jan Drchal, Ph.D. Supervisor’s signature	prof. Ing. Tomáš Svoboda, Ph.D. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt

Student’s signature

Acknowledgements

I thank to my supervisor, Jan Drchal, for the guidance through the Machine Learning and NLP methodology and for supporting my research interests. I would also like to thank to my family, who always supported me during my studies and throughout the process of working on this thesis.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague, May 16, 2022

Abstract

Automatic detection of bias in media represents a possible way to more objective and factual writing. This work focuses on media bias and addresses the problem of binary classification of media bias in the Czech news environment. The literature and methodology for different aspects of bias are reviewed. Then, a set of datasets related to media bias is collected and analyzed. Utilizing machine translation, eight parallel Czech datasets are presented, one of which being a large-scale dataset of 360k sentences. Additionally, a novel Czech dataset CWNC (Czech Wiki Neutrality Corpus) for bias detection with 5766 sentences is automatically extracted from Wikipedia. The following experiments then show the effects of pre-training on combinations of currently available datasets, suggesting a positive effect of pre-training on datasets focused on subjectivity. Czech BERT-based model with the best parameters and setting then achieves an F1 score of 80.2% on the selected target dataset. Finally, a trained classifier is utilized to classify real-world data from various news sources.

Keywords: natural language processing, bias detection, media bias,subjectivity detection, text classification

Supervisor: Ing. Jan Drchal, Ph.D

Abstrakt

Automatická detekce zaujatosti v médiích představuje možnou cestu k objektivnějšímu a faktičtějšímu psaní. Tato práce se zaměřuje na zaujatost médií a zabývá se problémem binární klasifikace zaujatosti médií v českém zpravodajském prostředí. Je provedena rešerše literatury a metodiky pro různé aspekty zaujatosti. Následně je shromážděn a analyzován soubor dat týkajících se zajuatosti médií. S využitím strojového překladu je prezentováno osm paralelních českých datasetů, přičemž jeden z nich je rozsáhlý dataset o 360 tisících větách. Kromě toho je z wikipedie automaticky extrahován nový český dataset CWNC (Czech Wiki Neutrality Corpus) pro detekci zaujatosti, s 5 766 větami. Následné experimenty pak ukazují vliv předtrénování na kombinacích aktuálně dostupných datasetů a naznačují pozitivní vliv předtrénování na datasetech zaměřených na subjektivitu. Český BERT model s nejlepšími parametry a nastavením pak dosahuje skóre F1 80.2% na vybraném testovacím datasetu. Nakonec je natrénovaný klasifikátor využit ke klasifikaci reálných dat z různých zpravodajských zdrojů.

Klíčová slova: zpracování přirozeného jazyka, detekce zaujatosti, mediální bias, detekce subjektivity, klasifikace textu

Contents

Figures

Tables

Chapter 1

Introduction

Consumption of media of all kinds increases as people become increasingly engaged in the online world. We usually learn about the events happening in the world through the online news media of our choice; therefore, it is essential that the media present fair, unbiased, and reliable information. Although there is a great deal of effort in pointing out potentially **biased** sources of information by humans[1], we still fail to classify the media in a broad and comprehensive manner. Automatic, machine classification of news texts may help humans on this journey toward a more reliable news environment.

Natural Language Processing (NLP) is a set of methods that work with text and aims to bring an understanding of textual data leveraging machine learning algorithms. Using the methods of NLP to classify news text has some limits for low-resource languages, such as the Czech language. In this work, I focus on exploring and applying current State-Of-The-Art (SOTA) methods on automatic detection of bias in Czech news.

1.1 Bias

Defining the word **bias** can be a bit tricky because, with different settings and different goals, the definition also changes. Much of the work done in Machine Learning (ML) research focused on **bias** also lacks a proper definition and often includes vague descriptions of its objectives [3].

In terms of ML, bias generally means a tilt, prejudice, or tendency that, during training, slips into the model and may subsequently lead to potentially unfair decisions. The bias is typically skewed toward some group of people, for example, based on their race, gender, etc.

For instance, an infamous example is Microsoft's AI chatbot that has picked up racist rhetoric from large racially biased data[2]. Another example is when large pre-trained language models exhibit stereotypical bias. Language models are often used to generate text, and such a biased model can generate harmful statements that contain social stereotypes [4].

Nowadays, these systems are used for decision-making in essential areas such as hiring, loans, and even justice. Therefore, the detection of this bias

[1] Websites like `allsides.com`, `mediabiasfactcheck.com` and `adfontesmedia.com`

[2] `https://futurism.com/delphi-ai-ethics-racist`

in ML models and subsequent mitigation of it have been widely studied [3]. This kind of ML bias research is often referred to as **unfairness in machine learning**.

However, as outlined before, besides the study of the models that reflect the biased nature of the data, one can focus on the origin of the bias introduced by humans in the first place. In other words, the writer's bias that is reflected in a text. This work focuses on the detection of bias in Czech **news**, so from now on I refer to bias as a **property of text** which can be potentially detected and classified.

1.1.1 Media Bias

The need to address bias in media articles arises from the ever-increasing social polarization. News exhibiting **Media Bias (MB)** can sway opinions and alter readers' beliefs. In this work, I refer to Allsides[3] definition[4] of the media bias:

> **Media Bias** - *noun.* The tendency of news media to report in a way that reinforces a viewpoint, worldview, preference, political ideology, corporate or financial interests, moral framework, or policy inclination, instead of reporting in an objective way (simply describing the facts). A media outlet may reveal bias in how it reports specific news stories or which stories they choose to cover, ie., deem more important than others to cover or emphasize.

An example of sentences from Allsides that exhibit explicit MB is as follows:

> *The World Health Organization is the world's best hope for fighting pandemics.*
>
> *Our leaders are cowards when we need them to be brave.*
>
> *Our justice system is a blight on our nation and makes a mockery of our ideals.*
>
> *The legislation never resulted in meaningful action.*

In these examples, there is a clear evidence of an author's opinion or state, on a particular problem imprinted into the statement. Although the definition is a bit too abstract, according to Allsides, MB can be decomposed into several features[5]. To name a few:

- **Sensationalism/Emotionalism** - Explicit sentiment in statement
- **Subjective Qualifying Adjectives** - Adjectives such as *extreme, awkward, serious,..*

[3] `https://www.allsides.com/` is a company that focuses on the non-automatic classification of news outlets with respect to their bias

[4] `https://www.allsides.com/blog/what-media-bias`

[5] `https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias`

- **Mudslinging/Ad Hominem** - Personal attacks, insulting, etc.

The diversity of these characteristics shows how complex and subtle the overall bias information can be. Therefore, a simple subjectivity or sentiment analysis would not be sufficient for cracking the MB detection.

Most of the features are of a lexical nature; on the other hand, there are other features that are practically not possible to detect automatically without a context, e.g., bias by **omitting information**, where it strongly depends on an outer context. In section 3.2 I refer to the family of these kinds of features as **informational bias**.

Previous examples show how statements that contain MB may be manipulative or persuasive to some extent. However, the presence of MB does not always imply malicious intent. It is in human nature to draw on experience, to express something, that we *believe* to be the truth, in a factual way. Thus, one can simply not be aware of their implicit bias. As the authors of Allsides suggest, the bias might even be desirable. For example, the *Commentary* format article often contains more bias (see 7.4.4), but its purpose is to present an opinion, and there is nothing wrong with that.

1.2 Outline and problem definition

In this work, my objective is to study the statement (sentence) level **binary classification** of the **media bias**, in the Czech language, using deep learning methods.

To do so, I break the task down into four research sub-tasks:

T1. Survey literature on automated media bias detection.

T2. Collect and inspect all datasets related to media bias.

T3. Investigate the possibilities of automatic creation of Czech datasets and create a unified collection of Czech datasets by translating the English ones.

T4. Experiment with collected datasets to train a Czech media bias classifier and evaluate the results on a target dataset.

And eventually aim to answer two research questions:

Q1. How well do the models trained on MB-related datasets generalize to the target dataset?

Q2. Does a pre-training on MB-related datasets help with media bias detection?

In addition, a trained classifier is applied to the Czech news corpora, and so the results of the real world MB classification are presented.

Chapter 2

Theoretical background

In this section, I briefly introduce SOTA methods and models used for classification in the experiment section 6.

Many of the problems in NLP are tasks of mapping one sequence to another; therefore, modern architectures were designed to tackle this problem efficiently.

2.1 Text representation

In order to process text, NLP models operate on the text on a **token** level. A token can be understood as the smallest unit of text and is a product of a process called **tokenization** (transforming text into a set of tokens).

A typical token unit is a word; however, tokens can be as small as a single byte. Currently, a standard way of tokenization is using WordPiece tokens, which is a balance between word-level and byte-level tokenization.

After tokenization, a numerical representation of the tokens is to be obtained. A naive way is to use a mapping for every word to an index in a predefined/obtained vocabulary. Such an approach suffers from the explicit ordering of the words, which may negatively influence the model. Another possible representation is **one-hot-encoding** where each word is represented by a vector that has 1 in a single row and zeros everywhere else. The size of such an encoding is proportional to the size of the vocabulary, making the feature space very sparse.

The current standard representations of words are word **embeddings**. Embeddings are fixed-size feature vectors. The dimension of the embeddings is a hyperparameter, usually with a value between 100 and 1000. Embeddings can be learned along a particular task in an Embedding layer. Or, it is possible to use available precomputed representations, such as word2vec or ELmo [5, 6].

2.2 Neural Networks

An Artificial Neural Network (ANN) is a model developed in the early 1940s [7]. The smallest unit of a Neural Network (NN) is a perceptron:

$$f(x) = \phi(w^T x + b) \tag{2.1}$$

Where **weight** vector w and **bias** term b are learnable parameters, x is an input feature vector and ϕ is an **activation function**. Activation functions are used for introducing non-linearities into the NN. In case of perceptron, it is a simple threshold function. Although the definition of Multi-Layer Perceptron (MLP) is loose, it can be understood as the simplest form of neural network with multiple connected perceptrons and a threshold activation function. In general, NNs usually use other activation functions such as sigmoid, ReLu, Tanh, LeakyReLu, etc.

2.3 Encoder-Decoder

Vanilla NN architecture operates on inputs of fixed length. For variable length input (for example, a sentence) Recurrent Neural Network (RNN) is used. Let $x = (x_1, x_2, ..., x_T)$ be the input sequence. RNN works on x sequentially, updating its **hidden state** vector h with some non-linear function, such as sigmoid, at each discrete time step. The final hidden state, also called **context vector** c_T, is preserved. Therefore, RNN is able to map the input of arbitrary length T to a fixed size vector c_T that captures the information of the entire sequence.

The sequence-to-sequence [8, 9] or **seq2seq** model aims to tackle the problem of mapping one sequence to another using two RNNs. The first RNN called **Encoder** is used to map the input sequence of arbitrary length to the fixed-size context vector. The context vector is then fed to a second RNN called **Decoder**. The decoder processes the context vector into a final sequence $y = (y_1, ..., y_k)$, by updating its hidden state vectors while generating outputs y_t.

There are several problems with this architecture. Firstly, when a long sequence is processed, the information from earlier parts of the sequence is "forgotten". Secondly, RNNs work in a sequential manner; hence, there is no room for parallelism. The **Transformer** architecture aims to solve these problems.

2.4 Transformers

Transformers [1] is a family of deep neural architectures that has revolutionized the field of NLP. The **attention** mechanism is a core principle behind the transformer architecture and solves some of the problems mentioned in the previous section.

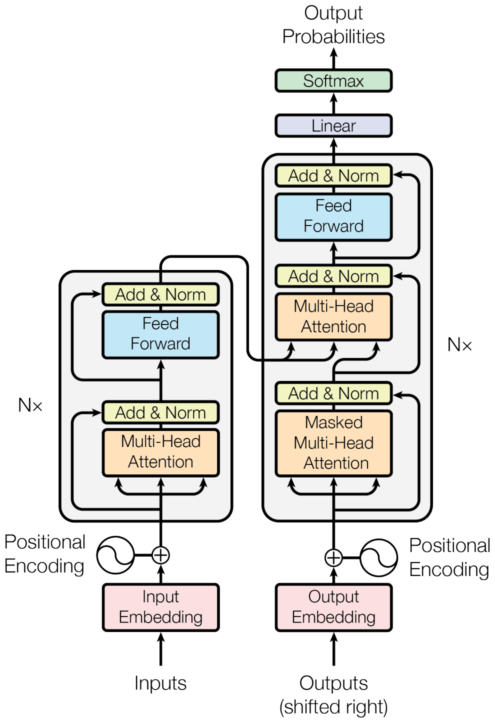


Figure 2.1: Transformer model architecture, reprinted from [1]

Current SOTA language models, such as GPT-3 [10], BERT [11], RoBERTa [12], all use the transformer architecture. Although their architecture can vary from the original one; for example, BERT only uses the Encoder module. Eventually, I have narrowed the choice of models to transformers rather than other neural architectures.

2.4.1 Attention

Attention [13, 14] is essentially a mechanism that allows the unit (whether it is on the decoder or the encoder side) to learn to focus on some segments of the input sequence more than others. This mechanism has been motivated by the problem of "forgetting" information in long sequences. This way, parts of the sequences that potentially drive the decision are represented more in hidden states than other less relevant parts.

2.4.2 Transformer architecture

Scheme of the original architecture can be seen in figure 2.1. The bottom of the architecture consists of an embedding layer enriched with **positional embeddings** that encodes the ordering information in the sequence.

Just as previous seq2seq models, the transformer also consists of Encoder and Decoder modules. Precisely, each module is a stack of n identical encoder or decoder layers. Each encoder layer consists of two main layers: the **self-attention** layer and the **feed-forward** neural network.

As particular words/tokens flow through the stack of encoders, the model gradually builds up their representation. The self-attention layer uses the attention mechanism to incorporate other words from the sequence into each particular word representation. The original transformer architecture performs this self-attention computation eight times, allowing the model to learn different relationships between words. This is what the term **Multi-Head** attention refers to. The resulting representations from these heads are then concatenated and reduced in dimension to obtain the final representation to be fed to the feed-forward neural network.

The Decoder modules work essentially the same, except it has an extra attentive layer that performs the attention computation over the output of the encoder stack. Since I only utilize the encoder part for more details, I refer the reader to the original transformer paper [1].

Although the computation of self-attention depends on other tokens, the forward pass of a feed-forward network can be done completely in parallel over the input sequence. Therefore, this design offers a significant computational speedup against the Encdoer-Decoder based on RNN.

2.5 Text classification

Text classification is a supervised learning task of assigning a particular text (word, sentence, or document) a category to which it belongs. A standard loss for classification is a Cross-Entropy loss. In the case of binary classification:

$$L_{BCE} = \frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \log \hat{y_i} + (1 - y_i) \cdot \log(1 - \hat{y_i})) \tag{2.2}$$

Where y_i denotes the ground-truth label, $\hat{y_i}$ the probability predicted by the model, and n is a number of samples.

2.5.1 Metrics

The most straight-forward way to evaluate the prediction ability of a classifier is to use the **accuracy** metric, which means counting correctly classified data.

$$accuracy = \frac{correct\ predictions}{total\ predictions} \tag{2.3}$$

This metric is feasible if the classes of the dataset are balanced. However, imagine a situation where 90% of the data belongs to one class and only 10% to another. The classifier, which always outputs the first class, achieves 90% accuracy even though its prediction capability is trivial. For unbalanced data, it is convenient to use the **F1** metric. The F1 score is a harmonic mean of *Precision* and *Recall*.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.4}$$

$$^1Precision = \frac{TP}{TP + FP} \tag{2.5}$$

Precision can be understood as "how precisely the model predicts a positive class", whereas recall

$$Recall = \frac{TP}{TP + FN} \tag{2.6}$$

can be understood as "how much of a positive class can model predict". The scores for each class are then averaged to obtain the final score. This is often referred to as *macro* averaging.

2.5.2 Transformers for text classification

The predictive power of transformers is behind many SOTA results, and text classification is no exception. For classification, usually only the encoder part of a transformer is utilized, although some define the classification problem as a sequence-to-sequence and incorporate the decoder too [15].

During tokenization, a special [CLS] token is prepended to a sentence. The token has its own embedding and flows through the stack of encoders just as any other token, with a difference that when the forward pass reaches the classification layer, only the [CLS] token is passed as an input. [CLS] token can therefore be understood as a sort of sentence embedding.

Usually, one or two dense layers with an activation function are sufficient as a classifier on top of the encoder stack. However, it is also possible to extract representations from any level of the encoder stack and run an arbitrary classification algorithm on top of it.

2.6 Transfer learning

Nowadays, the true power of transformers lies in **transfer learning** [16]. It is a process where some knowledge is not learned from scratch but transferred from a previously trained model. Since large transformer models such as BERT or RoBERTa have millions of parameters, it would be extremely costly to train them from scratch every time.

Such large models are usually pre-trained on an extensive corpus of data. There are several common **unsupervised** pre-training tasks that allow these models to learn contextual representations of words without supervision. For instance, Masked Language Modelling (MLM) is a task in which a random sample of tokens in the input sequence is replaced with a [MASK] token, and the model learns to predict the original token. Another unsupervised pre-training task is a Next Sentence Prediction (NSP) which is self-explanatory.

[1] TP,TN,FP,FN denotes to True positive, True Negative, Fasle Positive, False Negative respectively.

2.7 Fine-tuning for classification

Having a pre-trained model, one can then train a classification head on top of pre-trained representations. Although the process of fine-tuning is more often adopted.

In the context of this thesis, fine-tuning is essentially a process of adapting the pre-trained model, as well as the classifier, to the target dataset. Both the parameters of the language model, e.g., BERT, and the parameters of the classification head are updated jointly. The weights of the classification layer are trained from scratch; on the other hand, the pre-trained language model already contains meaningful linguistic features, so in the process, its parameters are fine-tuned for bias-specific representations.

2.8 Notes on Multi-Task learning

When talking about training a neural network, we usually talk about fitting a network to one particular task. However, fine-tuning language models on small datasets often results in overfitting. Multi-Task learning [17] serves as a good **generalization** technique and a potential solution to this problem.

In essence, Multi-Task Learning (MTL) means learning multiple tasks together within one model; thus, representations for all tasks are learned together. Tasks do not have to share loss functions; they can have their own task-specific heads on top of the shared model.

Chapter 3

Bias Detection

Before turning my attention to media bias, I have examined several other relevant bias detection topics. At the beginning of my research, I studied the possibilities of applying gender bias detection to Czech News. Therefore, the following small section is dedicated to my results and examination of one of the gender-focused datasets.

3.1 Gender bias detection

Most of the work on gender bias aims to study gender bias embedded in models and other methods to measure, clarify, and possibly mitigate it.

There is clear evidence that current language models possess implicit gender bias. Whether it means, in terms of learning biased embeddings [18], or simply underrepresentation of a particular gender in the data [19].

Yet, my work aspires to classify news texts; therefore, I examined the possibilities of gender classification in text.

I closely followed the approach of Dinan et al. [20]. They define three dimensions of gender bias: bias when speaking $ABOUT$ someone, TO someone or AS someone. Target classes are $\{$masculine,feminine,neutral$\}$.

The *bias* here simply means an aspect of the statement that implies the gender of a particular person along the dimensions. To make this definition more clear, for example, the authors further propose that an unbiased sentence would be a sentence in which a machine learning model would not be able to classify a gender because there would basically be no difference between the classes. Yet, in a real-world scenario, sentences **are** influenced by gender, and therefore such classification is possible.

To measure this kind of bias over all three dimensions, a large-scale dataset **md_gender**[1] has been collected. The authors train a transformer model (2.4) using MTL (2.8) paradigm, to capture all three dimensions. However, only the $ABOUT$ dimension and a very small fraction of the AS dimension are publicly available, so I focused only on the $ABOUT$ dimension.

[1] `https://huggingface.co/datasets/md_gender_bias`

- **md_gender** - is a collection of automatically labeled large-scale data gathered from various sources around the internet, where gender annotation of a particular dimension is provided (eg., gender information of a user in an internet discussion). It also includes a small gold-labeled dataset for evaluation with 785 data points for the *ABOUT* dimension.

3.1.1 Initial experiment

To transfer the results of the paper mentioned above to the Czech environment, I sampled 150k sentences from across all datasets with an *ABOUT* dimension label and translated them via **DeepL** machine translator (more on machine translation in section 5.1). Then I managed to train a RoBERTa-based model [12] that achieved an F1 score of 80% on the small gold-labeled evaluation dataset. Unfortunately, the results are not comparable to the original English experiments because I took a **single-task** approach and omitted other dimensions completely. I will share the trained model together with translated data on HuggingFace[2] hub, and I also present a demo. An example of the demo can be seen in Appendix A.

3.1.2 Discussion

The gender classifier, such as this one, can be used to determine what percentage of a particular article in the Czech news environment is about men, women, or is completely genderless. This statistical indicator could help to keep the writing more balanced or provide insight into already published writings.

Moreover, gender bias could also be an interesting task choice in the MTL setting, with respect to media bias. However, since this is just an initial experiment and is not further developed, I suggest that a clear-cut methodology and datasets review should be performed.

3.2 Media bias detection

When it comes to automatic detection of media bias, the standard is to use supervised learning. Most of the previous work in media bias used hand-crafted features together with traditional[3] ML algorithms. For example, Hube et al. [21] used a lexicon-based approach with various lexicons (sentiment, bias, subjective, and other linguistic features). Although hand-crafted feature-based approaches offer fairly reasonable explainability of the model's decision, they were outperformed by neural networks and have been mostly replaced by them.

The majority of current research focuses on **sentence-level** classification [22, 23, 24, 25], however, there has also been an effort to lift the classification to the **article-level**.

[2] `https://huggingface.co/`

[3] By traditional I refer to all ML models that are not deep neural networks.

Article-level classification is usually more difficult since it is inconvenient to put the whole article through the neural network. Even though such things as document embeddings exist [26], bottom-up solutions are usually used. A simple approach would be to classify all sentences and count the frequency. Eventually, I used this approach when applying the classifier to Czech news corpora in section 7.

However, additional high-level features such as the position of bias, frequency, or ordering, have been studied and proved to be effective in article-level classification [27, 28].

As I outlined in the previous section, MB can be divided into two classes, where one depends on the outer context, and the other does not. This is commonly referred to as **informational** and **lexical** bias. There have been efforts to classify informational bias with varying context sizes [29], although a majority of the work focuses on lexical bias, and I follow this standard as well.

Various pre-training and fine-tuning strategies have been studied regarding sentence classification. However, one of the most promising approaches is using an MTL to tackle the problem (see 2.8 for more details). Although there are already some results of applying MTL to the detection of MB [24, 30], empirical studies suggest that a large number of tasks should be used to allow MTL to shine [31]

Chapter 4

Datasets

Due to the complex nature of MB, different datasets try to capture different aspects of it. In this section, I present a collection of all datasets related to biased writing and subjectivity detection available. Later, this collection is used to study the proposed research questions Q1 and Q2 and to build a final classifier. For details see experiment section 6.

As stated above, this work focuses only on sentence-level classification; thus, data annotated on article-level were not considered.

I divided the available datasets into 3 main families:

- Subjectivity bias
- Wikipedia bias
- Media bias

Wikipedia bias could also be considered as a form of subjective bias, but all the Wiki data come from the same distribution[1]. and environment; hence I find it reasonable to put them together.

4.1 Subjectivity Datasets

Datasets that contain annotations of explicitly subjective expressions.

4.1.1 SUBJ

It is reasonable to include datasets that focus on the detection of subjectivity, as it is one of the MB characteristics. The Subjectivity dataset (SUBJ) [32] consists of 10000 sentences gathered from movie review sites. Sentences are labeled as subjective and objective with 1:1 ratio.

The data were collected automatically. The authors made an assumption that all reviews from Rottentomatoes[2] are subjective, and all IMBD plot summaries[3] are objective. For each class, 5k sentences were sampled randomly.

[1] Some datasets are different samples from the same larger corpora

[2] https://www.rottentomatoes.com/

[3] www.imdb.com

4.1.2 MPQA

Multi-**P**erspective **Q**uestion **A**nswering (MPQA) Opinion corpus is another dataset that can be used for subjectivity detection. I used the MPQA Opinion corpus version 2.0, which consists of 692 articles from 187 different news sources. In total 15802 sentences. All articles are from June 2001 to May 2002.

The corpus offers a rich annotation scheme [33] that focuses on sentiment and subjectivity annotations.

To extract the bias information, I focused on two types of annotations:

- Direct subjective
- Expressive subjective

These annotations were present if any form of subjectivity was suspected by the annotator. Each annotation consists of indices of span in the text and properties. For each sentence in the corpus, I extracted labels as follows:

If there was at least one annotation **direct_subjective** or **expressive_subjectivity** with span inside the sentence and the intensity tag was not *low*, the sentence was labeled as *subjective* ∼ *biased*. All other sentences were extracted as *objective* ∼ *unbiased*.

This approach has produced 9484 subjective sentences and 6318 objective sentences.

4.2 Media Bias datasets

Datasets that focus on media bias specifically.

4.2.1 BASIL

BASIL dataset [34] comprises 300 articles with 1727 sentence level bias annotations. The authors of the dataset distinguish between **lexical** and **informational** bias.

The annotations were performed by two experts and further resolution discussions have later led to 0.56 and 0.7 Inter-Annotator Agreement (IAA) score for lexical and informational bias, respectively.

Even though BASIL brings sufficient annotation quality, most of the labeling resulted in informational bias annotations, leaving only 478 sentences for the lexical bias class. Informational bias requires a different approach to detection [29] and usually depends dramatically on the context. Therefore, I extracted all sentences with the informational label as a neutral class.

4.2.2 Ukraine Crisis Dataset

This dataset [35] offers 2057 sentences with binary media bias labels. All sentences are related to one topic - the Ukraine-Russian crisis. Data were gathered from 90 news sources.

The authors introduce rich annotations for each sentence. Each of them looks at the bias from a different perspective, called *bias dimensions*:

1. Hidden Assumptions and Premises
2. Subjectivity
3. Framing

In addition, the *overall bias* annotation is presented. Together, the data include 44547 fine-grained annotations. For simplicity, I only included the overall bias annotation. Even though this dataset encompasses comprehensive bias information, it also suffers from a low IAA score. Specifically, Krippendorff's $\alpha = -0.05$ [36].

4.2.3 NFNJ

The NFNJ[4] dataset provides 966 sentences from 46 articles with annotations on a fine-grained level.

Authors share the dataset for research purposes; however, the public version differs from the one described in the original paper. Therefore, while extracting the final dataset, I made a few assumptions:

In the raw data, contributions from multiple annotators on each sentence are provided. Therefore, I extracted the labels as a simple arithmetical mean of the labels. Furthermore, the original labels stand for

- 1: 'neutral'
- 2: 'slightly biased but acceptable'
- 3: 'biased'
- 4: 'very biased'

To obtain the final labels in an unbiased-biased format, I simply assumed sentences with mean-score ≤ 2 as neutral and > 2 as biased.

The Fleiss Kappa IAA score averaged at zero, making it practically unusable as a standalone dataset.

4.2.4 BABE

Bias **A**nnotations **B**y **E**xperts (BABE)[38] is a key media bias dataset from Media Bias Group (MBG)[5], which is to the best of my knowledge, the highest quality media bias dataset to this day. It builds on top of MBIC [23] which is a smaller crowd-sourced dataset.

BABE contains 3700 sentences. 1700 sentences are from MBIC, which were extracted from 1000 news articles, and in addition extended by 2000

[4] [35] refers to this dataset as NFNJ, however, in the original paper [37] the name is not presented.

[5] `https://media-bias-research.org/`

Americans know President Donald Trump is an outrageous, scandal-ridden character.
Biden said he would seek Muslims to serve in his administration.
Biden's shift radically leftward reflects that of his party.
Anti-vaccine groups take dangerous online harassment into the real world.

Table 4.1: Example of biased and unbiased sentences from **BABE**. Red and green represent biased and unbiased annotations, respectively.

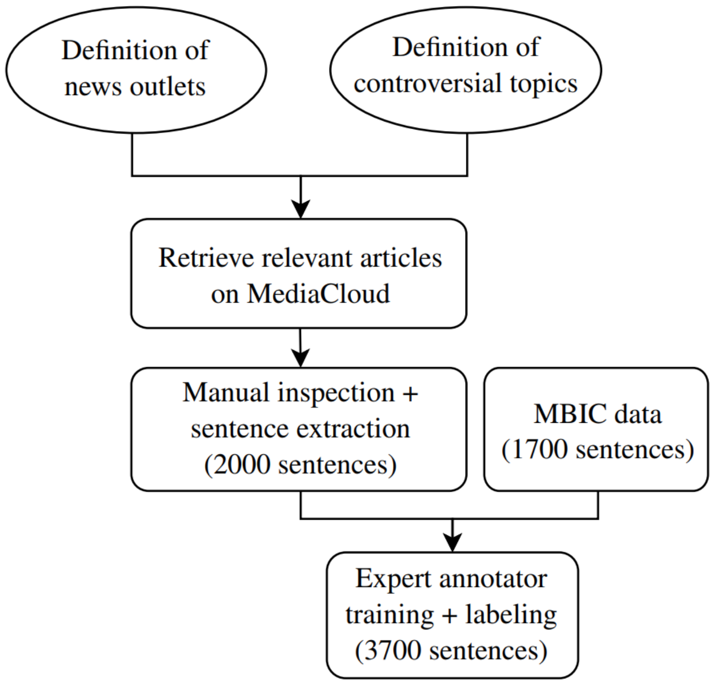

Definition of news outlets
Definition of controversial topics
Retrieve relevant articles on MediaCloud
Manual inspection + sentence extraction (2000 sentences)
MBIC data (1700 sentences)
Expert annotator training + labeling (3700 sentences)

Figure 4.1: Data collection and annotation pipeline of **BABE**, reprinted from [2]

more sentences, altogether covering 12 topics, annotated with binary bias indications. In addition, the annotations were enriched with a list of biased words. However, the presence of biased words does not always result in an overall biased sentence label. Examples of BABE data points can be seen in table 4.1.

It has been annotated by eight experts resulting in IAA Krippendorfs $\alpha = 0.39$, which exceeds other media bias datasets by a significant margin. It also provides detailed information about the annotator background, making it a **reliable** source of information. The pipeline of the collection of BABE can be seen in figure 4.1.

This dataset plays a pivotal role in my approach to media bias detection and is selected as a target for tuning and evaluating language models in chapter 6.

4.3 Wikipedia datasets

Due to annotation costs and the overall lack of large-scale datasets in the media bias setting, many researches [39, 40, 25] used Wikipedia's Neutral Point Of View (NPOV) policy[6] to construct a large-scale corpora **automatically**.

Wikipedia's NPOV policy is a set of rules that aim to preserve neutrality in Wikipedia articles. Some examples of NPOV principles are as follows:

- Avoid stating opinions as facts.
- Avoid stating facts as opinions.
- Prefer nonjudgmental language.

When neutrality is contested, a Wikipedia article can be moved to NPOV dispute by tagging it with {{NPOV}} or {{POV}}[7] template. Debate on specific details of neutrality violations is then initialized among editors and eventually resolved, leading to the removal of the tag.

This editorial information can be leveraged to extract parts of the text that violate the NPOV and their unbiased counterparts. However, it has been shown [25, 41] that such automatic extraction can suffer from noisy labeling. In some cases [25] up to 60% of data points from the positive class were actually neutral.

Even though these datasets introduce a large number of samples that are highly related to media bias, they are all sampled from Wikipedia's environment, which can be very different from the news environment.

4.3.1 Wiki Neutrality Corpus

Wiki Neutrality Corpus (WNC) [39] is a parallel corpus of 180k pairs of biased and unbiased sentences. For the collection of the data, 4.3 approach was adopted. The authors crawled revisions from 2014 to 2019. Each revision has been processed to check if it contains any variation of *POV* related tags inside. This approach yielded 180k pairs such that the sentence before edit is considered biased and the modified/added sentence after edit is considered neutral/unbiased.

In addition to WNC, 385k of sentences that have not been changed during the NPOV dispute were extracted as neutral and for word-level classification purposes, a subset of the WNC corpus, where only one word is changed in the biased-unbiased pair, were added.

4.3.2 CW-HARD

Hube et al. [25] constructed a dataset based on NPOV, where only revisions with one sentence diff were filtered. However, because of the potentially noisy

[6] `https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`

[7] Other POV related variations are often used.

Dataset	Size	Annotation	Agreement
SUBJ	10.000	automatic	-
MPQA	15.802	annotators	high
BASIL	1.727	annotators	medium
Ukraine Crisis Dataset	2.057	crowdsourcing	low
NFNJ	888	crowdsourcing	low
BABE	3673	annotators	medium
WNC	362.990	automatic	-
CW-hard	4953	crowdsourcing	low
WikiBias	8198	annotators	high

Table 4.2: Comparison of all bias related datasets collected

outcome, 5000 sentences were sampled and annotated using crowdsourcing. However, the Krippendorff's Alpha agreement score measured only $\alpha = 0.124$, which is generally considered low.

After filtering out sentences that annotators labeled with the "I don't know" option, the final dataset consists of 1843 statements labeled as biased and 3109 labeled as neutral, a total of 4953 sentences.

4.3.3 WikiBias

This is, to this date, the latest dataset based on Wikipedia. The authors [41] closely follow the approach of WNC [39] and extract another parallel wiki corpus of 214k sentences. To achieve a higher quality corpus, 4099 sentence pairs were randomly sampled and labeled by trained annotators. As a result, WikiBias-Manual dataset consists of 3400 biased and 4798 neutral sentences annotated with high IAA score of Cohen's $\kappa = 0.734$ [42].

4.4 Unused datasets

Some datasets focus on a slightly different task, yet still carry potentially useful information. Such data can be useful in a MTL setting (2.8). To name a few that are focused on the detection of ideology:

- **NewsB** - Consists of labels capturing the authors political ideology (liberal, conservative) Labeled through distant supervision.
- **IBC** - Also focuses on ideology detection; however, is not publicly available.

4.5 Summary

In this chapter, I introduced all resources that are potentially useful for media bias analysis and are publicly available. The overview of all datasets and its properties can be seen in figure 4.2.

BABE dataset is generally a good benchmark, and its translated parallel version will be used for evaluation in this work. Combining other listed datasets for pre-training is studied in 6. Unfortunately, many of the datasets suffer from noisy labeling and low IAA greatly. Therefore, its usability is considerably limited.

Chapter 5

Czech datasets

Despite the relatively sufficient number of datasets in English, there is essentially no suitable Czech dataset.

In essence, three options are feasible to solve this problem. The most promising way is to annotate a new gold-standard dataset. However, media bias is a non-trivial, complex, and subtle linguistic feature; hence, a lot of effort must be put into annotator training and eventually filtering of implicitly biased annotations.

Another way is to use an automatic approach. For example, Allsides[1] provide annotations on source and article-level with expert annotation quality. However, since I focus on a statement level only, using such data leads to oversimplification and results in a very noisy dataset. Regardless, it can still be used for domain-specific pretraining [2]. Unfortunately, there is no Czech site that would provide **useful** bias information on neither source nor article level. The server Nadační fond nezávislé žurnalistiky (NFNZ)[2] provides a scoring for different news sources. Yet, only a fraction of their scoring is related to the actual linguistic aspect of the writing. Most of the scoring is based on meta-information such as transparency, proper citation, advertisement, etc.

Nonetheless, the automatic creation of a dataset can be done in a convenient way, as described in section 4.3. Despite the limitation caused by the size of the particular Wikipedia, this approach is suitable for the Czech environment, as the Czech Wikipedia has a comparably large editor base[3] ranking #26 in a number of edits worldwide. I took this approach and present a **new parallel corpus** for bias detection based on Czech Wikipedia (5.4.2).

Finally, for low-resource languages, it is reasonable to translate English datasets. As one of my contributions to bias detection in Czech news, I reviewed, collected, and translated most of the relevant datasets described in chapter 4 using **DeepL**, and finally processed them into a unified format (5.2).

[1] https://www.allsides.com/unbiased-balanced-news

[2] https://www.nfnz.cz/

[3] https://en.wikipedia.org/wiki/List_of_Wikipedias

5.1 Machine Translation

Since the translation of large datasets by human translators would be too costly and, from a time perspective, practically impossible, automatic machine translation systems are used. In recent years, machine translation, like other fields of NLP, has experienced a significant increase in performance due to the rise of the attention mechanism and complex transformer architectures (2.4).

Modern machine translation models use the **encoder-decoder** where the encoder part distills (encodes) the information from the input sequence, and the decoder part is responsible for decoding this distilled information and mapping it to a sequence in the target language. For more details, see section 2.

For the translation of the datasets, **DeepL** translator, which is a purely[4] NMT based system, is used.

5.2 Processing

Every dataset has been processed into "sentence,label" format, where *label* $\in \{0, 1\}$ stands for **unbiased** and **biased**, respectively. Using this simplified data format makes merging and combining several datasets convenient. All cases have been preserved.

5.3 Translated data

All translated datasets are listed below. I hope that this collection will serve as a good starting point for future MB research in Czech News.

- BABE-CS
- Basil-CS
- WikiBias-CS
- CW-hard-CS
- MPQA-CS
- NFNJ-CS
- SUBJ-CS
- UA-crisis-CS
- WNC-large-CS[5]

[4] For example Google combines Neural Machine Translation (NMT) with statistical approaches, other systems incorporate hardcoded rules, etc.

[5] Additional *large* is added to distinguish between large translated WNC and the Czech version of WNC.

Together, approximately 400k bias-labeled translated sentences were collected. I will share the listed datasets on HuggingFace[6] hub. The distribution of the datasets can be seen in figure 5.1. The WNC is not included in the plot because it represents 87% of all data.

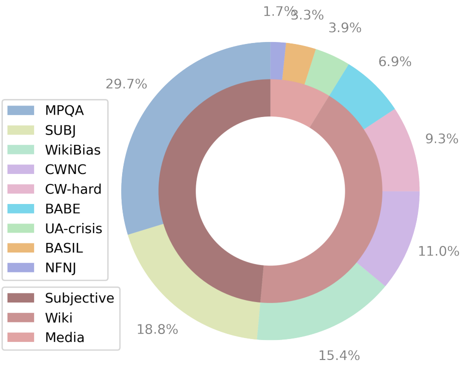


Figure 5.1: Dataset distribution in Czech collection of bias datasets (Without WNC)

5.4 Czech Wiki Neutrality Corpus

Finally, I present two novel parallel corpora extracted directly from Czech Wikipedia. To the best of my knowledge, these are the only original Czech datasets related to media bias detection. The only partially relevant dataset is **SubLex**[43] which is a subjectivity lexicon focusing mainly on sentiment. However, lexicon-based approaches are nowadays outperformed by neural models.

I followed two main existing approaches, both relying on the extraction of revisions that includes the {{NPOV}} tag or its variation. The NPOV tag also has its Czech version *Nezaujatý Úhel Pohledu (NÚP)*. However, the Czech version is practically not used, and so for the extraction, the English variations were used.

5.4.1 CWNC-noisy

I closely followed the [44] approach and used their publicly available script. Firstly, a file with all pages and its complete edit history is downloaded from the wiki dump[7]. I used the *20220201* version. Then, the pages with edits containing one of the NPOV-related tags are filtered, and the process of sentence extraction follows. This approach yielded 15k sentences; however, it uses a rather trivial assumption that when the NPOV tag is removed, **all** removed sentences are biased and all added are unbiased.

[6] https://huggingface.co/

[7] https://dumps.wikimedia.org/cswiki/

Nizozemsko je známé svým pokrokovým liberálním postojem vůči psychoaktivním drogám.
Nizozemsko je známé svým liberálním postojem vůči psychoaktivním drogám.
Mezi jeho nejznámější a zvlášť populární je jeho hudba ke hrám a filmům, která téměř zlidověla.
Mezi jeho nejznámější a zvlášť populární je jeho hudba k divadelním hrám a filmům, která v některých případech téměř zlidověla.

Table 5.1: Example of pairs of biased sentences and their rewritten neutral form

This annotating strategy led to a very noisy dataset, and for this reason, I excluded this dataset from further experiments entirely.

5.4.2 CWNC

This dataset was created following the [39] approach. The process is the same as described in section 4.3. I used *20220201* snapshot of Wikipedia, which was, at the time , the latest snapshot that included all the necessary files. I used the script publicly available on Github[8], with a few slight modifications so the processing fits the Czech language properties:

1. Used Regex was extended to exclude czech words that contain "pov" inside eg. povstání, povlak etc.
2. All cases has been preserved.
3. Czech Morphodita tokenizer was used[9]

The final dataset consists of:

- 3k of *before* and *after* sentence pairs
- 1.7k subset where only one word has been changed
- 7.5 sentences, where the change was rejected or reversed, implying neutrality of the original sentence.

In total, 5766 sentences. The neutral corpus, which contains only neutral sentences, is saved for a potential need of oversampling. Two examples of CWNC sentence pairs can be seen in figure 5.1

5.5 Not translated

Due to a large size of some datasets, I was unable to translate more than one large-scale dataset. For this reason, the NewsB dataset has not been translated.

[8] https://github.com/rpryzant/neutralizing-bias
[9] https://ufal.mff.cuni.cz/morphodita/users-manual

Chapter 6

Experiments

Finally, to complete the last research task proposed in the introduction T4, the datasets collected in section 4 are leveraged to build a **media bias classifier**.

To do so, a transformer model (2.4) is **fine-tuned** on the BABE dataset (4.2.4).

Furthermore, to answer the two research questions, Q1 and Q2 and to push the performance of the classifier, the effect of further pre-training[1] are studied.

A scheme of the whole process can be seen in figure 6.1. Firstly, a baseline fine-tuning is performed to select the suitable language model. Then a limited hyperparameter tuning of the model follows. Subsequently, the tuned model is then pre-trained on combinations of auxiliary datasets reviewed in section 4. The impact of the pre-training is studied through 1) direct evaluation on the BABE and 2) fine-tuning and evaluation on the BABE.

The optimal pre-training strategy and hyperparameters are then used to build the final clasifier.

6.1 Models

Architecture-wise, a classifier consisting of a dense[2] layer is attached to the pre-trained language model to perform the binary classification task. The tested language models were pre-trained either solely on Czech data (monolingual) or on multiple languages jointly (multilingual). The scheme of the particular architecture used can be seen in figure 6.2.

As opposed to English language, there is a relatively low number of Czech pre-trained language models available. A list and a brief summary of all the models tested can be found in the following:

- **RobeCzech** [45] - RoBERTa-based model with 125M parameters. Like its original counterpart, it is trained with the MLM task, on 4,917M tokens of Czech corpora.

[1] Primary pre-training is the original unsupervised pre-training task executed by the authors of the particular language model.

[2] Sometimes is referred to as fully connected layer.

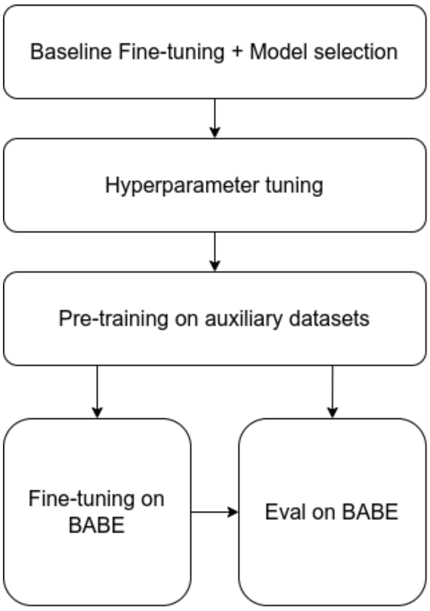

Baseline Fine-tuning + Model selection
Hyperparameter tuning
Pre-training on auxiliary datasets
Fine-tuning on BABE
Eval on BABE

Figure 6.1: Scheme of experiments.

- **Czert** [46] - BERT-based model with 110M parameters, trained with NSP tasks. All Together trained on 37GB of Czech text.
- **FERNET-C5** [47] - BERT-based model trained with the MLM and NSP task on 93GB of Czech text from the Common Crawl project.
- **FERNET-News** [47] - RoBERTa-based model trained with MLM task on 20GB of Czech News text.
- **SlavicBert** [48] - BERT-based model with 179M parameters, trained on four languages: Russian, Bulgarian, Czech, and Polish. The model is trained on all 4 languages at once. The model is not trained from scratch, but it is a fine-tuned version of mBERT.
- **mBERT** [11] - BERT-based model with 179M parameters trained on corpora of 104 languages, including Czech, with MLM task.

6.2 Experimental setup

All models are fetched, trained, and evaluated using the HuggingFace API[3]. The maximum sequence length is set to 128 tokens. All training parameters can be seen in the Appendix B.

A small portion (15%) of the target dataset is left aside as a **test set** at the beginning and is used only for the final evaluation to ensure that no test data leak into the training data.

[3] https://huggingface.co/docs

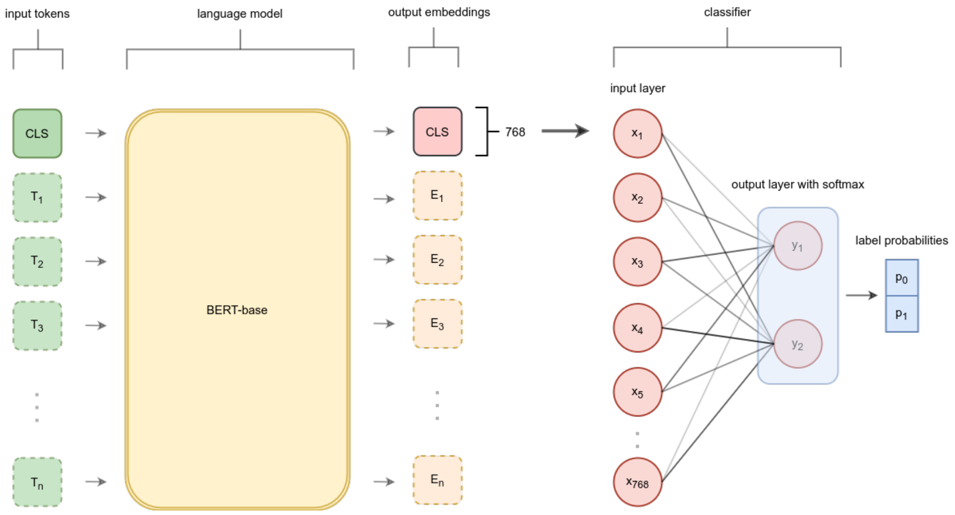


Figure 6.2: Scheme of a text classification architecture used in the fine-tuning of BERT-based models.

Every fine-tuning, except the one performed on the final test set, is evaluated using a 10-fold cross-validation (CV). This helps to get more realistic estimates of model performance than a simple train-validation split would give. The only evaluation metric used for all experiments is the F1 score with *macro* averaging[4].

All training has been done on the RCI[5] cluster node with 4 x NVIDIA Tesla V100 with 32GB GPU graphic memory.

6.3 Baseline

As a baseline, all Czech and multilingual models listed are fine-tuned on BABE and evaluated using a 10-fold stratified CV.

Because of the novelty of the CWNC, I also perform a baseline evaluation over this dataset, but later it is only used as an auxiliary dataset.

The hyperparameters used are the same as those used by the authors of BABE [23]. However, the authors used early stopping together with CV and used the validation split inside CV to early stop. This may lead to data leakage. Which can subsequently lead to too optimistic results.

A solution to this problem would require another split for validation, but at this point, the size of the training data is already shrunk significantly. Therefore, I did not use early stopping together with CV at all. Instead, I fixed the number of epochs to 3 as suggested by the authors of BERT [11] . All other hyperaparemeters remained unchanged; AdamW optimizer is used with an initial learning rate of 5e-5 and a batch size of 64.

The baseline evaluation of all Czech models used can be seen in table 6.1. The final F1 score is averaged across all folds.

[4] The F1 score is computed for both classes and averaged.

[5] `http://rci.cvut.cz/`

target\model	Czert	RobeCzech	mBERT	FERNET-C5	FERNET-News	SlavicBERT
BABE	0.776	0.774	0.733	**0.781**	0.566	0.754
CWNC	0.732	0.709	0.734	**0.747**	0.443	0.741
mean	0.754	0.742	0.734	**0.764**	0.505	0.748

Table 6.1: F1 scores of baseline fine-tuning. Best scores for each dataset are highlighted.

The model that performs best on the BABE is **FERNET-C5**. It also performs best on the novel CWNC dataset; therefore, it is a suitable candidate for further tuning. From now on, all experiments are performed using this model.

6.4 Hyperparameter tuning

I restricted the search space of hyperparameters only to the combinations of:

- **Batch size** $\in \{16, 32\}$
- **Learning rate** $\in$ {2e-5,3e-5,5e-5}
- **Epochs** $\in \{2, 3, 4\}$

As the authors of the original BERT paper suggest [11]. Then I ran a grid search with CV. The overall best parameters were as follows:

{learning_rate = 3e-5, batch_size = 32, epochs=3}

The model with the best parameters achieved a 0.784 F1 score (∼0.4% improvement against baseline).

6.5 Combining Datasets

This section is dedicated to the study of the influence of pre-training on combinations of datasets. Trying all combinations would result in training 511 models[6], which is infeasible. Therefore, I decided to experiment with pre-training on five subsets of datasets with regard to their bias information, to see which of them can serve as a good initialization for fine-tuning on BABE. The combination sets are as follows:

- **SUBJe** - is a combination of the SUBJ and MPQA dataset, both of which focus on explicit subjective bias.
- **MB** - is a combination of NJNJ, UA-crisis and BASIL dataset which are all from the MB family.

[6] Given a set of n elements, number of subsets is 2^n. Here, we have a set of nine datasets, resulting in 512 subsets. 511 without an empty set.

	BABE	SUBJ	WIKI	MB	WNC	ALL
Pre-trained + Fine-tuned	0.7835	0.7875	0.7797	0.7702	0.7825	**0.7878**
Pre-trained	-	0.5542	0.6344	0.4631	**0.6697**	0.6423

Table 6.2: F1 scores of pre-trained models with and without further fine-tuning.

- **WIKI** - are all datasets collected from Wikipedia. It consists of CW-hard, WikiBias, and CWNC. The three datasets were collected automatically with respect to NPOV violations as described in 4.3.

- **ALL** - This one is simply a combination of all datasets except the WNC.

- **WNC** - WNC is almost 90% of all data; therefore, I perform experiments on this dataset separately.

In every combination, the data were randomly mixed and subsequently downsampled, so that the classes were balanced. For each combination, 20% of data were used as a validation set to decide the optimal number of epochs for pre-training. The convergence of validation losses can be seen in the figure 6.3. The number of epochs for pretraining were chosen as follows: 1, 3, 1, 2, 1 for SUBJe, MB, WIKI, ALL and WNC respectively.

This procedure yields five pre-trained models for further experiments.

6.5.1 Pre-training + Evaluating

To answer the question Q1, the pre-trained models from the previous section are evaluated on BABE without any fine-tuning on it. This way, it can be studied how well each model trained on each set can transfer knowledge to the detection of MB in BABE. Thus, possibly unveils the relatedness to BABE. The results can be seen in table 6.2. In the table, this pre-training **without** further fine-tuning is referred to as **Pre-trained**.

Models pre-trained on Wikipedia data, both **WIKI** and **WNC** perform relatively well compared to **MB** and **SUBJ**. This suggests that the bias distribution in WIKI datasets is the closest to BABE.

The best performing model was pre-trained on the largest dataset, WNC. The model achieved an F1 score of 0.67 on the target dataset which is 21% more than the model pre-trained on MB set

During pre-training, the F1 score of the WIKI and the MB set both peaked around 70% on their validation sets; however, the model trained on MB generalized very poorly to BABE data as opposed to the WIKI model (0.46 against 0.63). I suspect that the low quality of the MB set and high size imbalance between the two sets played an important role in this result.

In conclusion, models that generalized to BABE best were pre-trained on Wikipedia datasets.

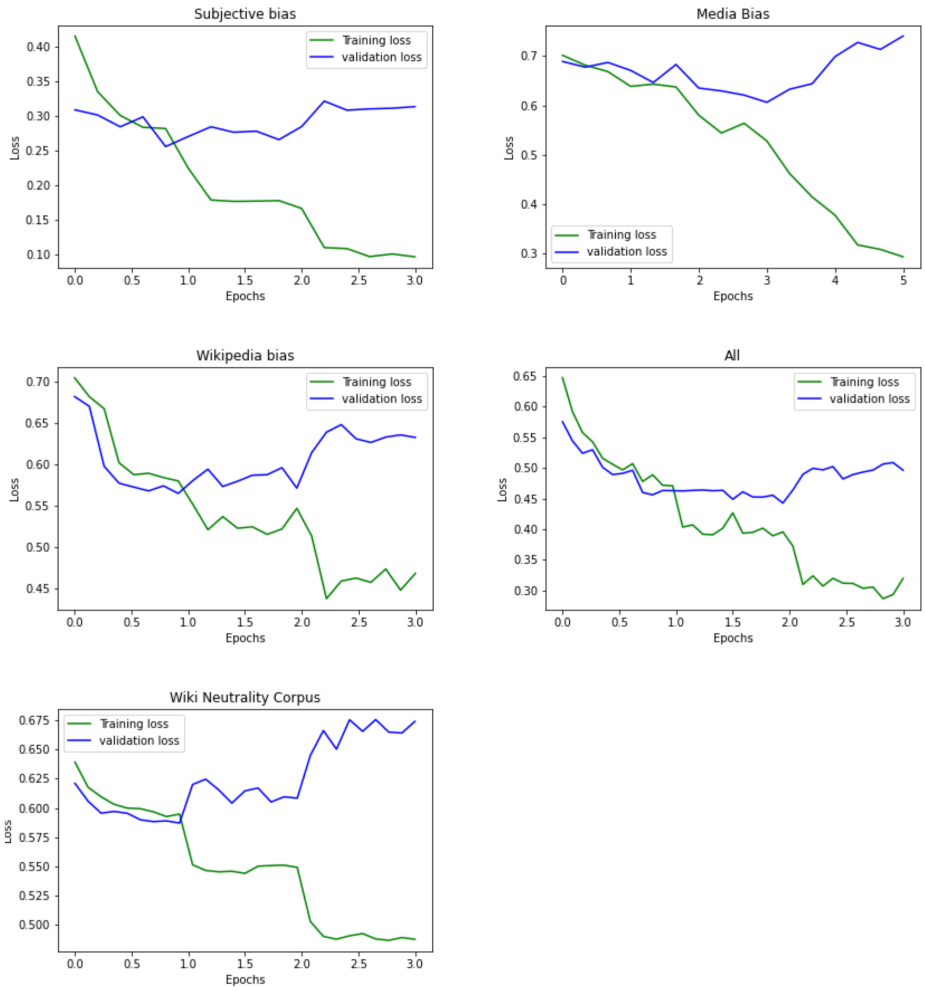


Figure 6.3: Convergence of validation loss of different dataset combinations

6.5.2 Pre-training + Fine-tuning

Secondly, pre-trained models are used as a sort of weight initialization for subsequent **fine-tuning** on the BABE. The results can be seen in table 6.2. This process is referred to as **Pre-trained + Fine-tuned**.

Fine-tuning a model pre-trained on **ALL** datasets combined resulted in the best performance; however, virtually the same performance was achieved by the model trained on the **SUBJe** combination. Importantly, the SUBJ split represents almost half of all data (see 5.1). Therefore, I assume that the performance of ALL model is high because of the presence of SUBJe in the training data.

Using the rest of the pre-trained models for fine-tuning actually hurt the performance. Yet, the difference is very small. The lowest score is achieved by fine-tuning the MB model. I suspect this is mainly due to the small size (2500 sentences) of the balanced MB set that may have led to overfitting.

In conclusion, using models pre-trained on the subjective datasets improved the performance of fine-tuning on BABE, but the increase was very small.

6.6 Final evaluation

The final FERNET-C5 model has been pre-trained with **ALL** datasets combination and fine-tuned with the optimal parameters (6.4) on BABE target dataset. Finally, on a **test** set it achieved an F1 score of **0.804**. A confusion matrix of predictions on the test set can be seen below 6.4

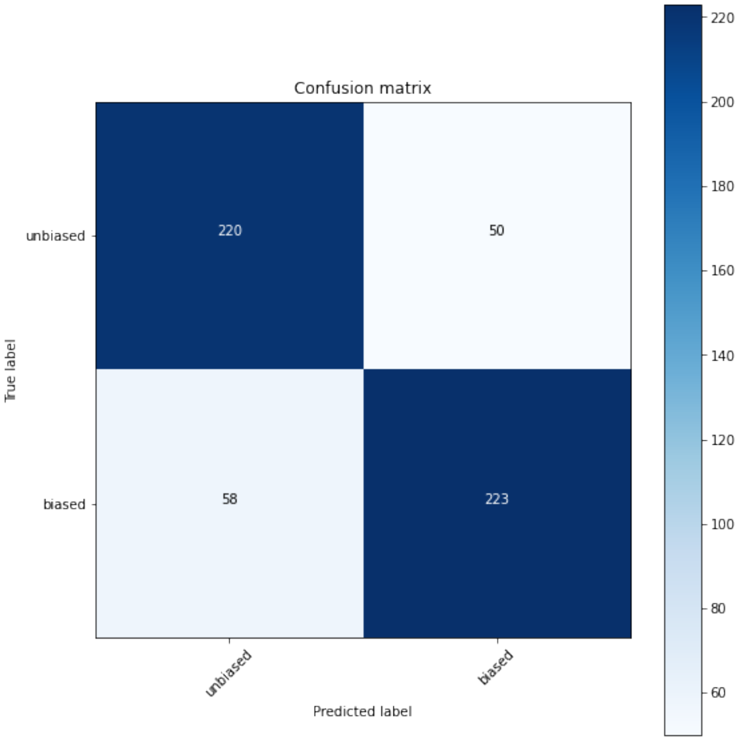


Figure 6.4: Confusion matrix of predictions on the test set.

For the final model that I share[7] on HuggingFace and which is used for inference experiments, the entire BABE dataset, including the test set, was used for training.

6.7 Discussion

The results suggest that subjectivity bias as opposed to media bias appears to be a bit more explicit and straightforward, since pre-training on the subjectivity task helped with MB detection, but proved insufficient without further fine-tuning (only 55% on BABE). This supports the assumption that MB is composed of many more superficial linguistic features (1.1.1).

[7] https://huggingface.co/horychtom/czech_media_bias_classifier

Also, despite the relatively high performance, the final score on the test set is not representative due to its size. The test set consists of ~ 500 sentences and therefore, may not adequately represent all bias information. For a better evaluation, I propose using a nested CV [49].

The authors of the original paper that introduces BABE [38] report an F1 score of ~ 0.8 for fine-tuned transformer model. That is approximately 2% higher score than I achieved on the Czech version. This may be caused either by the noise introduced into the data during machine translation or by the possibly lower quality of the Czech pre-trained models.

Perhaps, a complete study with more models could be performed, but that would require an enormous number of trained models. Essentially, these results show that there was a minimal gain over the baseline (**+0.7%**).

The results indicate that the overall low quality and limited size of the available datasets make their use for media bias detection impractical.

Chapter 7
Inference on Czech News

Finally, the classifier built in the previous section is used to classify media bias in the Czech news corpus. The results across different domains, sections, and granularities are presented.

7.1 Data

For this purpose, the **SumeCzech**[1] dataset has been used [50]. It is originally a dataset meant for training summarization models; however, it consists of complete news articles in JSON format and therefore is suitable.

The dataset includes five Czech news domains: novinky.cz, idnes.cz, ceskenoviny.cz, denik.cz, lidovky.cz.

A list of the fields available in the dataset can be seen in the following:

- `text` : body of the article
- `headline` : headline of the article
- `abstract` : abstract of the article
- `subdomain` : domains with some additional information about the topics, e.g. lidovky.cz has sport.lidovky.cz for sport news.
- `section` : a topic of the article
- `published` : date of publication
- `length` : number of characters in the text

7.2 Pre-processing

The SumeCzech contains around one million articles. For the experiments, I only used the validation and test split, which comprises approximately 90k articles. Furthermore, only data from the ceskenoviny.cz domain were oversampled from the large train set to match the size of other domains.

[1] `https://ufal.mff.cuni.cz/sumeczech`

All subdomains were stripped such that only a domain remained. The prefix of the domain was additionally imputed as a section. See an example in the following:

```
subdomain:= novinky.cz <- sport.novinky.cz
section:= sport <- sport.novinky.cz
```

Also, I decided to exclude all blogs. Therefore, all the data that contained the 'blog' substring in its subdomain were removed.

To balance the domains, I sampled 4000 articles from each domain, resulting in a final dataset consisting of 20000 articles.

7.3 Methodology

For each article in the dataset, the following procedure was executed:

1. Text and abstract are splitted into sentences using `sent_tokenize` nltk function[2].

2. Each sentence from the text, abstract and a headline is classified with binary label 0,1 using the media bias classifier (6.6).

3. Each sentence from the text is assigned with 'reported_speech' indicator if matched the regular expression for quoting extraction.

4. The percentage of biased sentences is calculated for text and abstract.

5. The percentage of reported speech among the sentences is calculated for the text.

This procedure results in the introduction of four new fields:

- `text_bias` : percentage of biased sentences in the text
- `abstract_bias` : percentage of biased sentences in the abstract
- `headline_bias` : bias label of the headline
- `quoting_ratio` : percentage of reported speech in the text

These fields are further used to study the nature of the media bias in the corpus. For data processing and statistics, third-party libraries were used, in particular, **numpy**[3], **pandas**[4], **scipy**[5] for data processing and **seaborn**[6] for visualizations.

[2] https://www.nltk.org/
[3] https://numpy.org/
[4] https://pandas.pydata.org/
[5] https://scipy.org/
[6] https://seaborn.pydata.org/

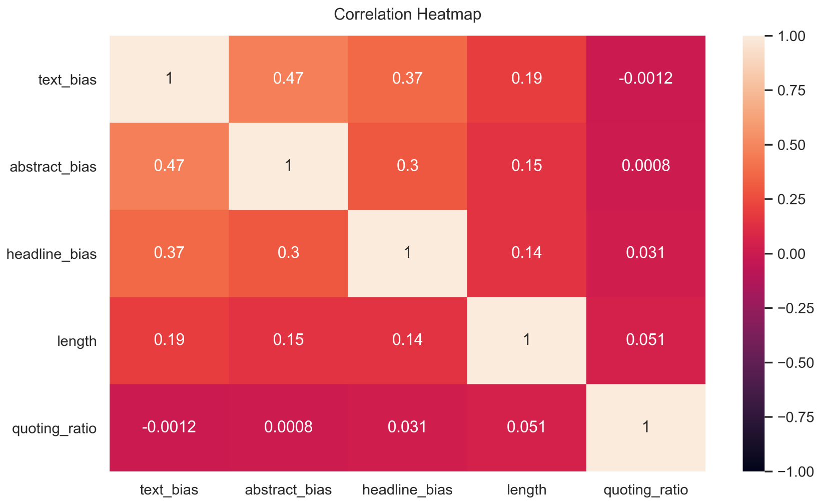


Figure 7.1: Correlation of dataset fields.

7.4 Results

In this section I present statistics of media bias across domains and topics and, additionally, its progression over time for the denik.cz domain.

7.4.1 Correlations

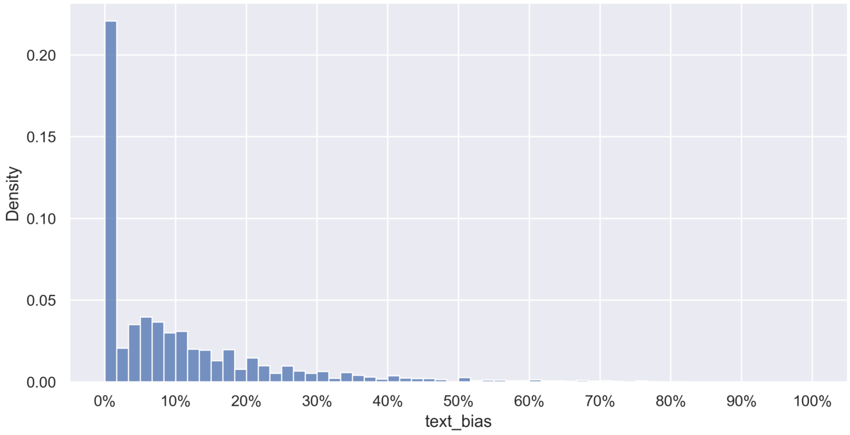
Firstly, a simple correlation heatmap is shown in figure 7.1. Bias of the abstract and the headline has a relatively high correlation with the `text_bias`. Both the abstract and the headline are a sort of an aggregate of the text; therefore, these results are justified.

Furthermore, there is some evidence that longer articles tend to be more biased, but the coefficient of correlation is relatively low.

Seemingly, there is no relationship between the `quoting_ratio` (frequency of reported speech) and the bias of the article.

7.4.2 Bias distribution

Histogram of the `text_bias` values can be seen in figure 7.2. In the first plot, we see that more than 20% of the classified articles had no biased sentences at all. To have a more detailed look at the distribution, I omitted the unbiased articles in the second histogram. Apparently, most of the biased articles have between 5-10% of biased sentences.



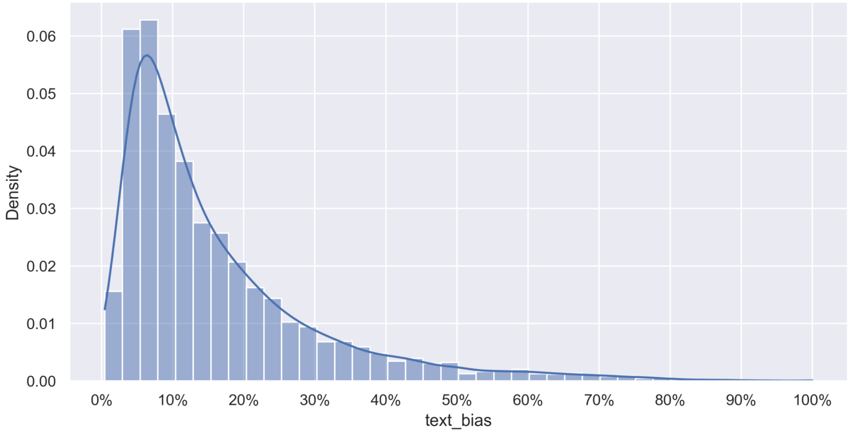


Figure 7.2: Distribution of text bias values over the dataset. Second plot is without the articles with 0 text bias.

7.4.3 Bias between domains

To study the bias distribution between domains, the data was grouped by domains and their `text_bias` and `headline_bias` within the domain were averaged. The bar plot of the results can be seen in figure 7.3. **ceskenoviny.cz** exhibit significantly lower average `text_bias` as well as the ratio of biased headlines (average `headline_bias`).

Furthermore, the difference between the bias distribution of the least biased and the most biased domain is presented in the figure 7.4. An x-axis is log-scaled for better visualization.

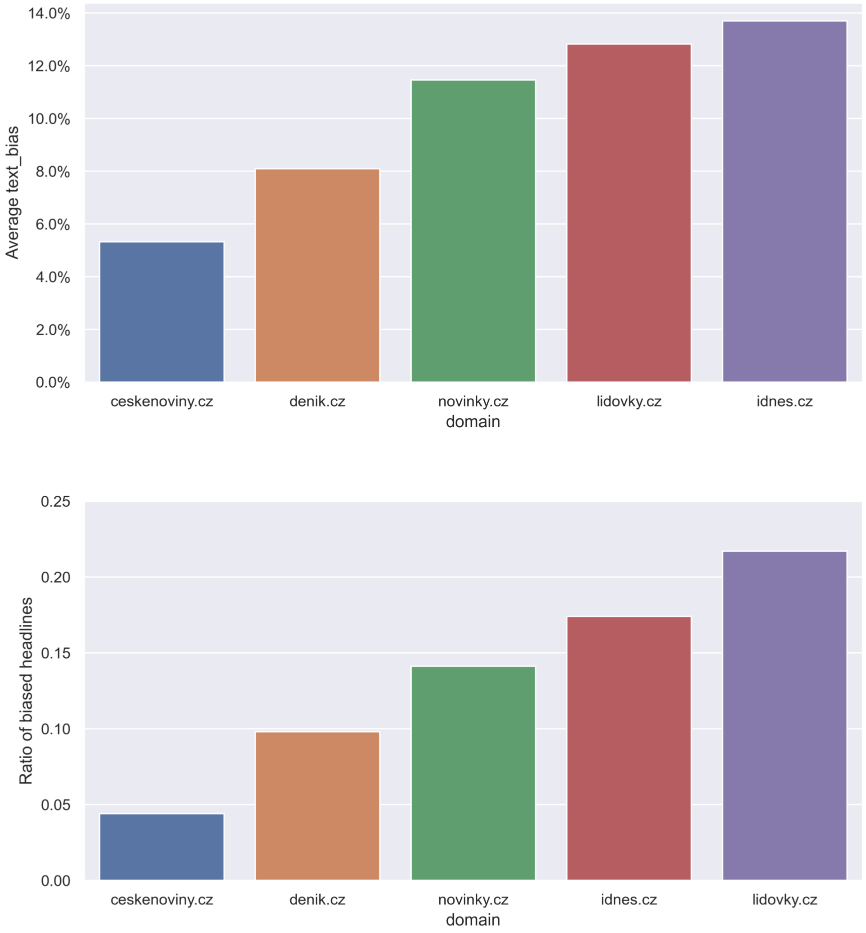


Figure 7.3: Comparison of average text bias and headline bias across the domains.

7.4.4 Bias between sections

In this experiment, data were grouped by sections (topics of the article) and their `text_bias` was averaged. I made an arbitrary choice to exclude sections that had less than 50 data points within the dataset, because such sections were way too specific.

The ten least biased and most biased sections[7] can be seen in figure 7.5. The values are in percents.

Sections that exhibit low average `text_bias`, for example *career*, *economics*, *business* are all of a rather factual, emotionless nature. On the other hand, sections with high average bias, such as *art*, *music*, and *culture*, are more of a subjective, personal nature.

[7] *rungo* is a fitness lifestyle section, *xman* is a male oriented section and *revue* is sort of a yellow press.

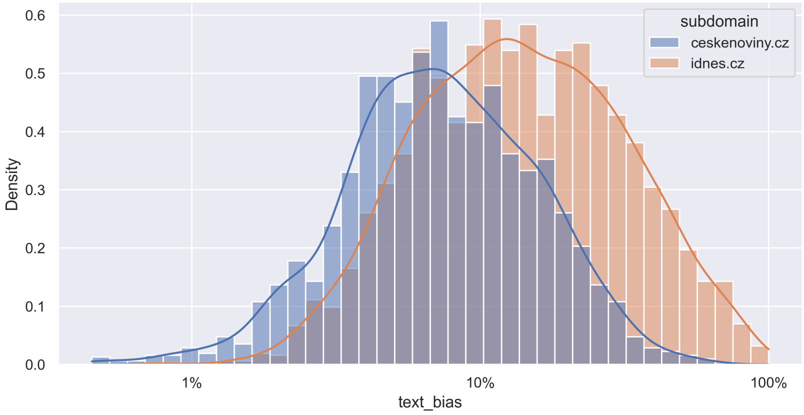
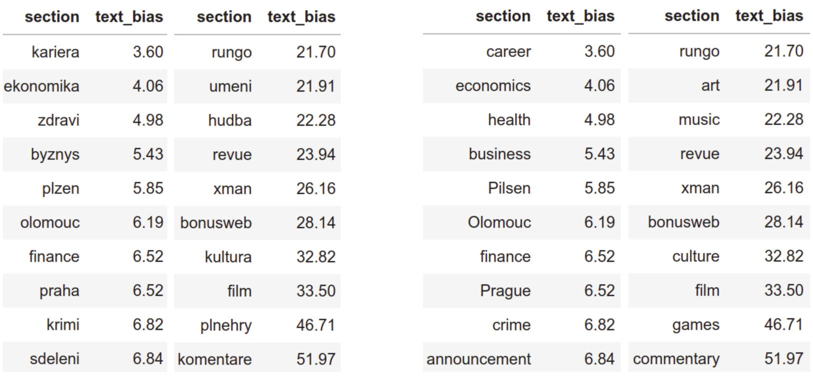


Figure 7.4: Distributions of bias between two domains.



section	text_bias	section	text_bias
kariera	3.60	rungo	21.70
ekonomika	4.06	umeni	21.91
zdravi	4.98	hudba	22.28
byznys	5.43	revue	23.94
plzen	5.85	xman	26.16
olomouc	6.19	bonusweb	28.14
finance	6.52	kultura	32.82
praha	6.52	film	33.50
krimi	6.82	plnehry	46.71
sdeleni	6.84	komentare	51.97

(a) : in Czech

section	text_bias	section	text_bias
career	3.60	rungo	21.70
economics	4.06	art	21.91
health	4.98	music	22.28
business	5.43	revue	23.94
Pilsen	5.85	xman	26.16
Olomouc	6.19	bonusweb	28.14
finance	6.52	culture	32.82
Prague	6.52	film	33.50
crime	6.82	games	46.71
announcement	6.84	commentary	51.97

(b) : in English

Figure 7.5: Ten least and ten most biased sections

Interestingly, the *commentary* section, exhibits by far the highest average `text_bias` (51.97%). This was an expected result, as the essence of the *commentary* format is to present an opinion, often subjective.

7.4.5 Denik.cz experiments

Finally, a progression of the media bias over time is examined. Because a distribution of the domains across the years was highly imbalanced, I decided to examine only the denik.cz domain.

To make the results as clear as possible, I only used data from one section. denik.cz data include articles from the years 2007 to 2016. For each year, 350 articles were randomly sampled and the `text_bias` was averaged. A graph of the progression can be seen in figure 7.6. I used a linear regression model to fit a line through the data points to highlight the trend.

Additionally, the same statistics, but with respect to months, is shown in figure 7.7. A blue stripe represents a confidence interval.

The experiment has shown a decreasing trend in media bias over a period of ten years. The inspection of this result and its possible causes is beyond the scope of this thesis and beyond the scope of the subject of media bias classification.

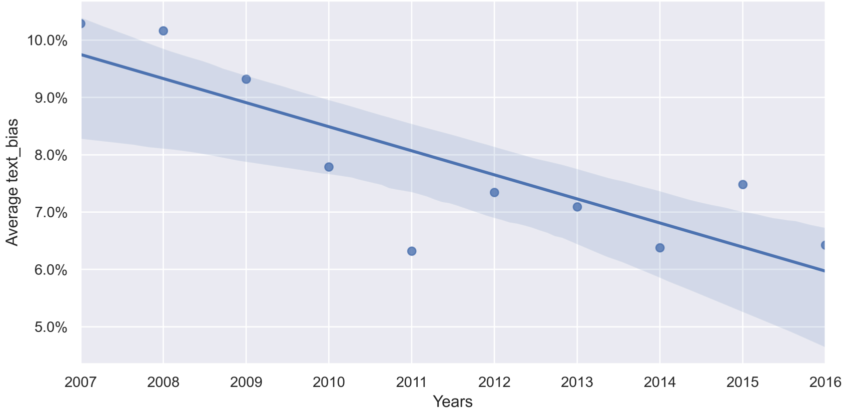


Figure 7.6: Progression of text bias of denik.cz over ten years.

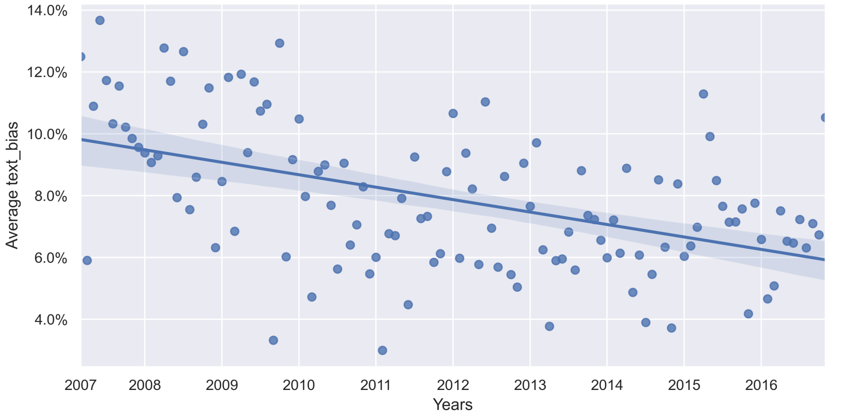


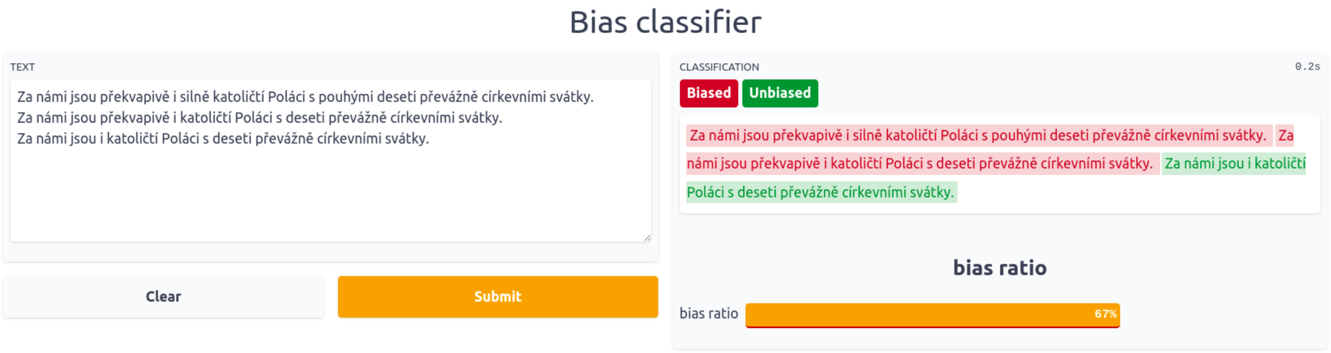
Figure 7.7: Progression of text bias of denik.cz over ten years. The bias is averaged over months.

7.5 Discussion

Even though these inference experiments have shown some interesting properties of the news, the data has some minor issues in this context.

For example, not all commentary articles are marked as commentary in the data. Also quoting of reported speech is sometimes missing.

On top of it, the classifier has around 80% accuracy (6.6), therefore these results should be taken with a grain of salt.



Bias classifier

TEXT

Za námi jsou překvapivě i silně katoličtí Poláci s pouhými deseti převážně církevními svátky.
Za námi jsou překvapivě i katoličtí Poláci s deseti převážně církevními svátky.
Za námi jsou i katoličtí Poláci s deseti převážně církevními svátky.

CLASSIFICATION 0.2s

Za námi jsou překvapivě i silně katoličtí Poláci s pouhými deseti převážně církevními svátky. Za námi jsou překvapivě i katoličtí Poláci s deseti převážně církevními svátky. Za námi jsou i katoličtí Poláci s deseti převážně církevními svátky.

bias ratio

bias ratio 67%

Figure 7.8: Example of the bias classifier demo usage.

7.6 Application

Additionally, I provide a simple web demo application for the reader to experiment with[8]. The app runs on HuggingFace's spaces[9] which is a free hosting service for demonstrating ML applications. For the frontend, Gradio[10] was used.

The user can insert arbitrary text in Czech language (text in other languages will result in meaningless outcomes). The text is then split into sentences and classified individually.

The application shows the results in two output windows; the first one *classification* displays inserted text with highlighted labels. The second one, *bias ratio*, displays the percentage of biased sentences in the text.

An example can be seen in figure 7.8. One example of the classification of the whole article can be seen in Appendix C.

[8] https://huggingface.co/spaces/horychtom/czech_media_bias_detection

[9] https://huggingface.co/spaces

[10] https://gradio.app/

Chapter 8
Conclusion

In this work, I collected and analyzed the literature and resources to study state-of-the-art media bias detection and also performed a minor experiment focused on gender bias detection. I presented a new Czech parallel dataset derived from Wikipedia with 5 765 sentences and, in addition, nine parallel translated Czech datasets to tackle the detection of media bias in Czech language, one of them large-scale (WNC with 360k sentences).

I trained and tuned the BERT-based FERNET-C5 language model for binary classification and achieved an F1 score of 0.804 on a small test subset of the BABE media bias dataset. I performed experiments on combining different datasets for pre-training the model to push the performance on the validation set. The pre-training on all datasets combined performed the best; however, both hyperparameter tuning and pre-training had generally a very low effect on performance, approximately +0.7% gain over the baseline.

Finally, the final classifier has been used to build a publicly available demo and to analyze a sample of articles from the SumeCzech dataset. The results of this study showed a trend in the progression of media bias over time and revealed a positive correlation between the headline bias and the average bias of the article.

8.1 Ethical Concerns

Although the performance of the current classifier is quite appealing, a standalone F1 score might not provide the appropriate evaluation of the ability of the model. Perhaps a human evaluation should also play a role in the process.

Also, the model's decisions are not easily clarifiable. The problem of explainability of the model is especially important when such classifier is brought to real-world applications.

8.2 Future perspective

As outlined in the introduction (1.1.1), according to allsides.com, media bias appears to be a combination of several potentially independent features, such as sentiment, agression, or subjectivity. In this thesis, this hypothesis has also been somewhat supported by the result that pre-training on the subjectivity task had the best influence on the detection of MB but a single-task model for subjectivity detection eventually performed worse than others. Therefore, as Spinde et al. [30] suggest, the multi-task approach could be used to improve the classification ability of the current classifier. Therefore, a future collaboration with MBG has been established and the MTL approach will be thoroughly studied.

Nevertheless, the current Czech classifier relies heavily on the translated datasets. I suggest that, for future improvement, a construction of an original gold-standard Czech dataset is essential.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[2] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa, “Neural media bias detection using distant supervision with babe - bias annotations by experts,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, (Dominican Republic), 2021.

[3] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of “bias” in nlp,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

[4] M. Nadeem, A. Bethke, and S. Reddy, “Stereoset: Measuring stereotypical bias in pretrained language models,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, 2021.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[7] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[14] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation,"

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[17] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[18] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[19] J. Sun and N. Peng, "Men are elected, women are married: Events gender bias on Wikipedia," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (Online), pp. 350–360, Association for Computational Linguistics, Aug. 2021.

[20] E. Dinan, A. Fan, L. Wu, J. Weston, D. Kiela, and A. Williams, "Multi-dimensional gender bias classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 314–331, 2020.

[21] C. Hube and B. Fetahu, "Detecting biased statements in wikipedia," in *Companion proceedings of the the web conference 2018*, pp. 1779–1786, 2018.

[22] M. Sinha and T. Dasgupta, "Determining subjective bias in text through linguistically informed transformer based multi-task network," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3418–3422, 2021.

[23] T. Spinde, L. Rudnitckaia, S. Kanishka, F. Hamborg, Bela, Gipp, and K. Donnay, "Mbic – a media bias annotation dataset including annotator characteristics," in *Proceedings of the iConference 2021*, (Beijing, China (Virtual Event)), 2021.

[24] N. Lee, B. Z. Li, S. Wang, P. Fung, H. Ma, W.-t. Yih, and M. Khabsa, "On unifying misinformation detection," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5479–5485, 2021.

[25] C. Hube and B. Fetahu, "Neural based statement classification for biased language," in *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 195–203, 2019.

[26] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," in *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 78–86, 2016.

[27] W.-F. Chen, K. Al Khatib, B. Stein, and H. Wachsmuth, "Detecting media bias in news articles using gaussian bias distributions," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4290–4300, 2020.

[28] W.-F. Chen, K. Al Khatib, H. Wachsmuth, and B. Stein, "Analyzing political bias and unfairness in news articles at different levels of granularity," in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, (Online), pp. 149–154, Association for Computational Linguistics, Nov. 2020.

[29] E. van den Berg and K. Markert, "Context in informational bias detection," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6315–6326, 2020.

[30] T. Spinde, J. Krieger, T. Ruas, J. Mitrović, F. Götz-Hahn, A. Aizawa, B. Gipp, and E. T.-b. M. Learning, "Exploiting transformer-based multitask learning for the detection of media bias in news articles,"

[31] V. Aribandi, Y. Tay, T. Schuster, J. Rao, H. S. Zheng, S. Vaibhav Mehta, H. Zhuang, V. Q. Tran, D. Bahri, J. Ni, *et al.*, "Ext5: Towards extreme multi-task scaling for transfer learning," *arXiv e-prints*, pp. arXiv–2111, 2021.

[32] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the ACL*, 2004.

[33] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2, pp. 165–210, 2005.

[34] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang, "In plain sight: Media bias through the lens of factual reporting," 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)*, pp. 6343–6349, 2019.

[35] M. Färber, V. Burkard, A. Jatowt, and S. Lim, "A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3007–3014, 2020.

[36] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.

[37] S. Lim, A. Jatowt, and M. Yoshikawa, "Understanding characteristics of biased sentences in news articles.," in *CIKM workshops*, 2018.

[38] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa, "Neural media bias detection using distant supervision with babe-bias annotations by experts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1166–1177, 2021.

[39] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang, "Automatically neutralizing subjective bias in text," in *Proceedings of the aaai conference on artificial intelligence*, vol. 34, pp. 480–489, 2020.

[40] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1650–1659, 2013.

[41] Y. Zhong, J. Yang, W. Xu, and D. Yang, "WIKIBIAS: Detecting multi-span subjective biases in language," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, (Punta Cana, Dominican Republic), pp. 1799–1814, Association for Computational Linguistics, Nov. 2021.

[42] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[43] K. Veselovská and O. Bojar, “Czech SubLex 1.0,” 2013. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[44] D. Aleksandrova, F. Lareau, and P. A. Ménard, “Multilingual sentence-level bias detection in wikipedia,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 42–51, 2019.

[45] M. Straka, J. Náplava, J. Straková, and D. Samuel, “Robeczech: Czech roberta, a monolingual contextualized language representation model,”

[46] J. Sido, O. Pražák, P. Přibáň, J. Pašek, M. Seják, and M. Konopík, “Czert – Czech BERT-like model for language representation,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, (Held Online), pp. 1326–1338, INCOMA Ltd., Sept. 2021.

[47] J. Lehečka and J. Švec, “Comparison of czech transformers on text classification tasks,” in *International Conference on Statistical Language and Speech Processing*, pp. 27–37, Springer, 2021.

[48] M. Arkhipov, M. Trofimova, Y. Kuratov, and A. Sorokin, “Tuning multilingual transformers for named entity recognition on slavic languages,” in *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, pp. 89–93, 2019.

[49] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[50] M. Straka, N. Mediankin, T. Kocmi, Z. Žabokrtskỳ, V. Hudeček, and J. Hajic, “Sumeczech: Large czech news-based summarization dataset,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Glossary

ANN Artificial Neural Network. 5

CV cross-validation. 26, 27, 31

IAA Inter-Annotator Agreement. 14–16, 18, 19

MB Media Bias. 2, 3, 12, 13, 21, 27, 28, 30, 41

MBG Media Bias Group. 15, 41

ML Machine Learning. 1, 2, 11, 39

MLM Masked Language Modelling. 8, 24, 25

MLP Multi-Layer Perceptron. 5

MTL Multi-Task Learning. 9–12, 18, 41

NFNZ Nadační fond nezávislé žurnalistiky. 20

NLP Natural Language Processing. 1, 5, 21

NMT Neural Machine Translation. 21

NN Neural Network. 5

NPOV Neutral Point Of View. 17

NSP Next Sentence Prediction. 8, 25

NÚP Nezaujatý Úhel Pohledu. 22

RNN Recurrent Neural Network. 5, 7

SOTA State-Of-The-Art. 1, 4, 6, 8

WNC Wiki Neutrality Corpus. 17

Appendix A

Gender classification results

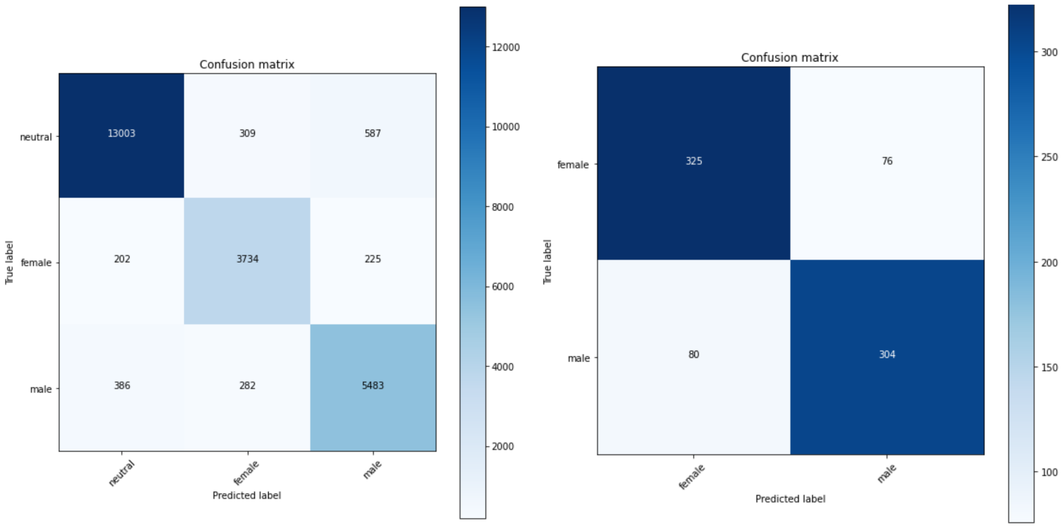
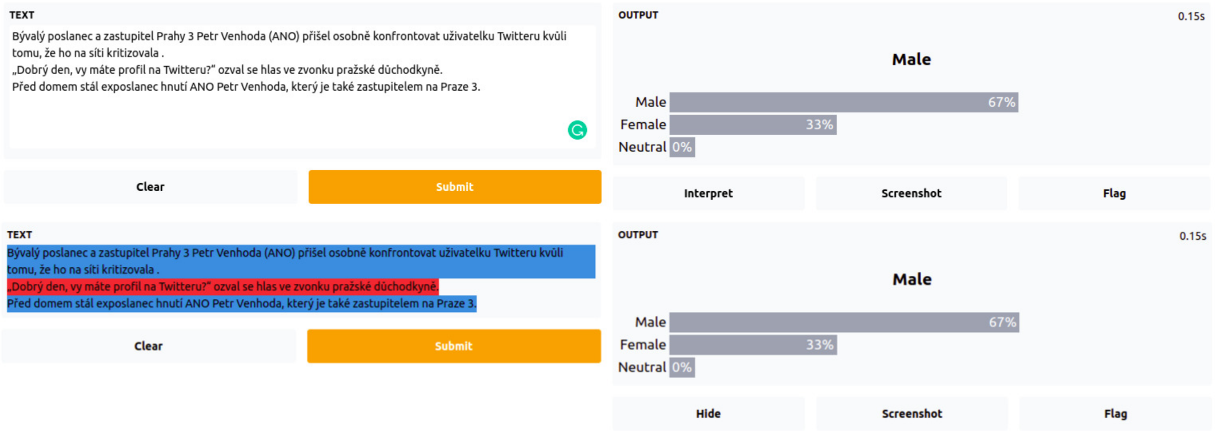


Figure A.1: Confusion matrices of gender classifier on test sets. On the left on large scale validation dataset. On the right on a gold-label small test set.



Bývalý poslanec a zastupitel Prahy 3 Petr Venhoda (ANO) přišel osobně konfrontovat uživatelku Twitteru kvůli tomu, že ho na síti kritizovala .
„Dobrý den, vy máte profil na Twitteru?“ ozval se hlas ve zvonku pražské důchodkyně.
Před domem stál exposlanec hnutí ANO Petr Venhoda, který je také zastupitelem na Praze 3.

Male

Male	67%
Female	33%
Neutral	0%

Bývalý poslanec a zastupitel Prahy 3 Petr Venhoda (ANO) přišel osobně konfrontovat uživatelku Twitteru kvůli tomu, že ho na síti kritizovala .
„Dobrý den, vy máte profil na Twitteru?“ ozval se hlas ve zvonku pražské důchodkyně.
Před domem stál exposlanec hnutí ANO Petr Venhoda, který je také zastupitelem na Praze 3.

Male

Male	67%
Female	33%
Neutral	0%

Figure A.2: Example of the gender bias classification.

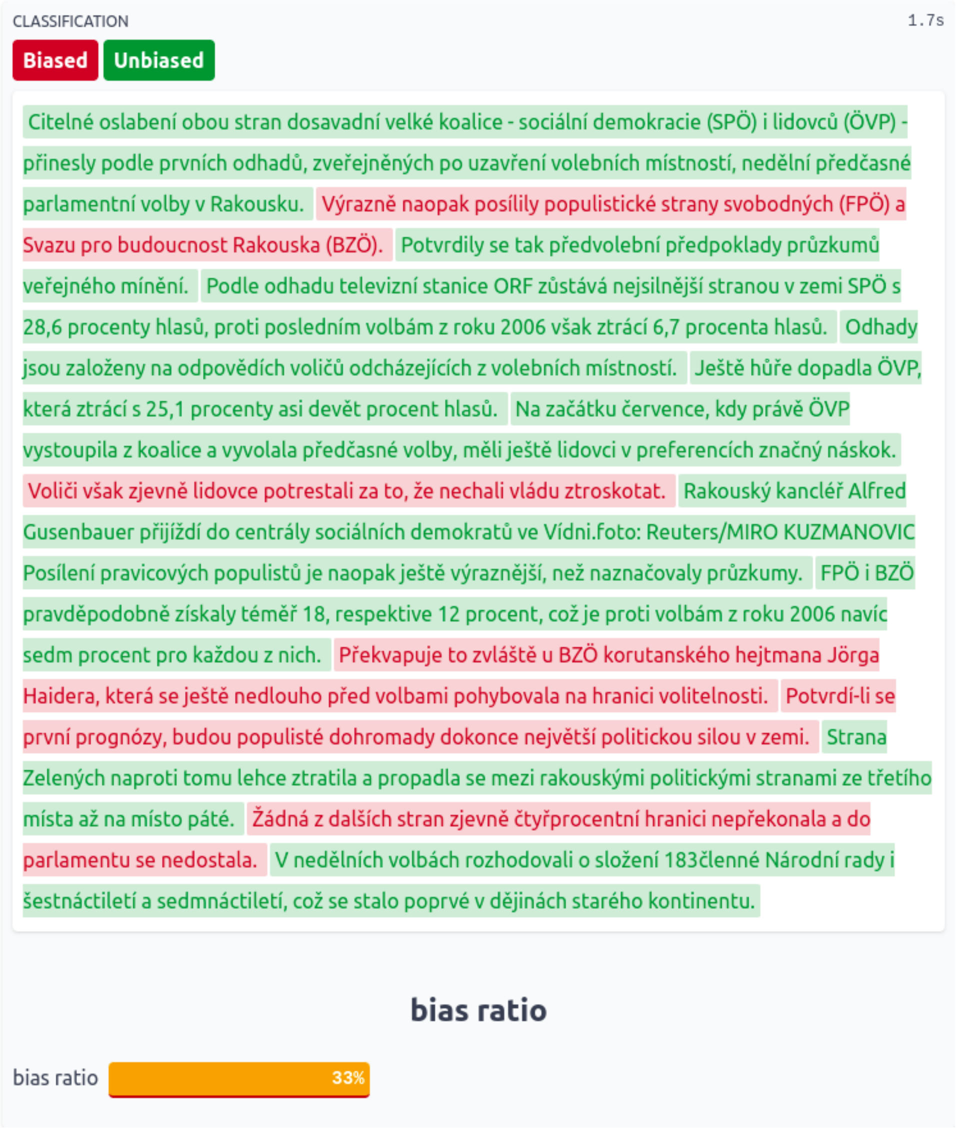
Appendix B

Training parameters

- per_device_train_batch_size: 32
- gradient_accumulation_steps: 1
- learning_rate: 3e-05
- weight_decay: 0.1
- adam_beta1: 0.9
- adam_beta2: 0.999
- adam_epsilon: 1e-08
- max_grad_norm: 1.0,
- num_train_epochs: 3,
- lr_scheduler_type": "linear"
- warmup_ratio: 0.0
- seed: 42
- no_cuda: false
- fp16: false
- remove_unused_columns: true
- load_best_model_at_end: false
- ignore_data_skip": false
- label_smoothing_factor: 0.0
- adafactor: false

Appendix C

Classified news article



Citelné oslabení obou stran dosavadní velké koalice - sociální demokracie (SPÖ) i lidovců (ÖVP) - přinesly podle prvních odhadů, zveřejněných po uzavření volebních místností, nedělní předčasné parlamentní volby v Rakousku. Výrazně naopak posílily populistické strany svobodných (FPÖ) a Svazu pro budoucnost Rakouska (BZÖ). Potvrdily se tak předvolební předpoklady průzkumů veřejného mínění. Podle odhadu televizní stanice ORF zůstává nejsilnější stranou v zemi SPÖ s 28,6 procenty hlasů, proti posledním volbám z roku 2006 však ztrácí 6,7 procenta hlasů. Odhady jsou založeny na odpovědích voličů odcházejících z volebních místností. Ještě hůře dopadla ÖVP, která ztrácí s 25,1 procenty asi devět procent hlasů. Na začátku července, kdy právě ÖVP vystoupila z koalice a vyvolala předčasné volby, měli ještě lidovci v preferencích značný náskok. Voliči však zjevně lidovce potrestali za to, že nechali vládu ztroskotat. Rakouský kancléř Alfred Gusenbauer přijíždí do centrály sociálních demokratů ve Vídni.foto: Reuters/MIRO KUZMANOVIC Posílení pravicových populistů je naopak ještě výraznější, než naznačovaly průzkumy. FPÖ i BZÖ pravděpodobně získaly téměř 18, respektive 12 procent, což je proti volbám z roku 2006 navíc sedm procent pro každou z nich. Překvapuje to zvláště u BZÖ korutanského hejtmana Jörga Haidera, která se ještě nedlouho před volbami pohybovala na hranici volitelnosti. Potvrdí-li se první prognózy, budou populisté dohromady dokonce největší politickou silou v zemi. Strana Zelených naproti tomu lehce ztratila a propadla se mezi rakouskými politickými stranami ze třetího místa až na místo páté. Žádná z dalších stran zjevně čtyřprocentní hranici nepřekonala a do parlamentu se nedostala. V nedělních volbách rozhodovali o složení 183členné Národní rady i šestnáctiletí a sedmnáctiletí, což se stalo poprvé v dějinách starého kontinentu.

Figure C.1: Example of classified article.

Appendix D

List of attachments

The attached thesis zipfile and data zipfile are structured as follows:

```
attachment1
├── data
│   ├── CS
│   ├── EN
│   └── Inference
├── demo
├── notebooks
│   ├── gender
│   └── media
├── src
├── README.md
└── thesis.zip

attachment2
└── inference_data.zip
```