Bachelor Project

Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Bias Detection in Czech News

Tomáš Horych

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

Assignment valid until: **30.09.2023**

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

We thank the CTU in Prague for being a very good *alma mater*.

Declaration

I declare that this work is all my own work and I have cited all sources I have used in the bibliography.

Prague, February 10, 2022

Prohlašuji, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškerou použitou literaturu.

V Praze, 10. února 2022

Abstract

This manual shows how to use the ctuthesis LaTeX class, what are the requirements, etc.

Keywords: bias detection, transformers, text classification

Supervisor: Ing. Jan Drchal, Ph.D

?abstractname?

Tento mánuál představuje LaTeXovou třídu ctuthesis, její použití, požadavky na systém atd.

Keywords: bias detekce, transformers, text classification

Contents


Figures
Tables

Chapter 1
Introduction

This is an introduction to my thesis, motivation. taky něco o nlp

1.1 Outline and Motivation

In this work, I focus on automatic binary classification of the presence of **media bias** in statements (sentences), using state-of-the-art language models and applying the classifier to Czech News. I gather all feasible resources for the Czech language, training, Czech models, and evaluation on Czech news corpora. I hope that my work will kickstart the research on media bias in the Czech environment, and so I present several future research proposals.

Before turning all 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, I dedicate a small section **??** to my results, an examination of one gender-focused dataset, and potential use cases.

1.2 Bias

Defining the word **bias** can be a bit tricky, because with different settings and different goals the definition also shifts. Much of the work done with bias also lacks a proper definition and often includes vague descriptions of its objectives [**?**].

In terms of Machine Learning (ML), bias usually means a tilt, prejudice, or tendency that, during training, enters the model and may subsequently lead to potentially unfair decisions. The bias is typically skewed towards some group of people, for example **racial bias**, **gender bias**, etc.

To put things into perspective, an infamous example is when Microsoft AI chatbot has picked up racist rhetoric from large racially biased data[1]. 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 may generate harmful statements that contain social stereotypes [**?**].

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

Nowadays, these systems are used for decision making in essential areas such as in hiring, loans, and even justice. Therefore, the detection of potential **unfairness** of ML models and subsequent mitigation of such biases have been widely studied [**?**].

However, 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 the human in the first place. Whether it is the presence of gender, stereotypical, or subjective bias, this kind of biased writing, especially in the news, can have a significant influence on people who consume it.

1.2.1 Media Bias

The need to address bias in the media arises from the ever increasing social polarization. News that exhibit **media bias** can sway opinions and alter readers beliefs. In this work I refer to Allsides[2] definition[3] 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.

Such bias can be decomposed into several features[4]. To name a few:

- **Sensationalism/Emotionalism** - Explicit sentiment in statement
- **Subjective Qualifying Adjectives** - Adjectives such as *extreme, awkward, serious,..*
- **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 is not sufficient. 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 or would require different approaches, e.g. bias by **ommiting information**, where it strongly depends on an outer context. In section **??** I refer to the family of these kinds of features as **informational bias**.

However, the presence of media bias does not always imply malicious intent. It is in human nature to draw on experience; thus, one can simply not be aware of their implicit bias. As the authors of Allsides suggest, it might even be desirable. For example, the *Commentary* format article often contains more bias, but its purpose is to present an opinion, and there is nothing wrong

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

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

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

with that. Trouble begins when the reader is not aware of it and anticipates or assumes objective journalism.

Chapter 2

State of the art

2.1 Gender bias detection

Most of the work done 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 learned, biased embeddings [**?**] or simply underrepresentation of a particular gender in the data [**?**].

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. [**?**]. They define three dimensions of gender bias: bias when speaking *ABOUT* someone, *TO* someone or *AS* someone. Target classes are {masculine,feminine,neutral}.

The word **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. Authors train a transformer model using Multi-Task Learning (MTL), to capture all three dimensions, however, only *ABOUT* dimension and very small fraction of *AS* dimension are publicly available, so I only focused on the first one.

- **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 data set for evaluation with 785 data points for the *ABOUT* dimension.

To transfer the results of the paper mentioned above to the Czech environment, I sampled 150k sentences from across all datasets with an *ABOUT*

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

dimension label and translated them via **DeepL** machine translator (more on machine translation in section **??**). Then I managed to train a RoBERTa based model that achieved an F1 score of 80% on the small gold labeled evaluation dataset. Unfortunately, the results are not comparable because I took a **single-task** approach and omitted other dimensions completely. I share the trained model together with translated data on HugginFace[2] hub and I also present a demo. An example of the demo can be seen in the appendix.

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 give an insight into already published writing.

2.2 Media bias detection

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

The majority of current research focuses on **sentence level** classification [**?**, **?**, **?**, **?**], however, the classification can be extrapolated to **article level**.

Article-level classification is usually more difficult, since it is quite problematic to put the whole article through the neural network, even though such things as document embeddings exist, bottom-up solutions are usually used. A naive approach would be to classify all sentences and simply count the frequency. Additional high-level features (eg., position of bias) have been studied and have been shown to be effective [**?**, **?**].

As I outlined in the previous section, media bias 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 [**?**], although it is a rather unique approach, and so neither I will focus on informational bias in this work.

Various pre-training and fine-tuning strategies have been studied; however, one of the most promising is using an MTL to tackle the problem. Even though there are already some results of applying MTL to the detection of media bias [**?**, **?**], empirical studies suggest that a large number of tasks has to be used to allow MTL truly shine. See **??** for more details about MTL.

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

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

Chapter 3

Datasets

Due to the complex nature of bias, 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 and later leveraged their bias information to augment smaller ground truth datasets. For details see experiment section **??**.

As stated before, this work only focuses on sentence level classification, thus article level data were not considered. I divided the datasets into 3 main families:

- Subjectivity bias
- Wikipedia bias
- Media bias

Wikipedia bias is also 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.

3.1 Subjectivity Datasets

3.1.1 SUBJ

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

The data were collected in an automatic way. The authors made an assumption that all reviews from Rottentomatoes[2] are subjective and all plot summaries from IMBD[3] are objective. Thus, the labels can be assumed to be noisy. For each class, 5k sentences were sampled **randomly**.

[1] some are even different samples from same larger corpora
[2] https://www.rottentomatoes.com/
[3] www.imdb.com

3.1.2 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 summing up to 15802 sentences. All articles are from June 2001 to May 2002.

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

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

- Direct subjective
- Expressive subjective

Which 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 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 labelled as *subjective* $\sim$ *biased*. All other sentences were extracted as *objective* $\sim$ *unbiased*.

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

3.2 Media Bias datasets

3.2.1 BASIL

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

3.2.2 Ukraine Crisis Dataset

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

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

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

In addition, the *overall bias* annotation is presented. Together, the data involve 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 low IAA score. Specifically Krippendorff's $\alpha = -0.05$.

3.2.3 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 truth labels in a 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, which makes it practically unusable as a standalone dataset.

3.2.4 BABE

Bias **A**nnotations **B**y **E**xperts (BABE) 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 [**?**] which is a smaller crowdsourced dataset.

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

It has been annotated by 8 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 **??**.

This dataset plays a pivotal role in my approach to media bias detection and is selected as a target for tuning language models in chapter **??**. Examples of BABE data points can be seen in **??**

[4] [**?**] refer to this dataset as NFNJ, however in the original paper the name is not presented.

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

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

Table 3.1: Example of biased and unbiased sentences from **BABE**

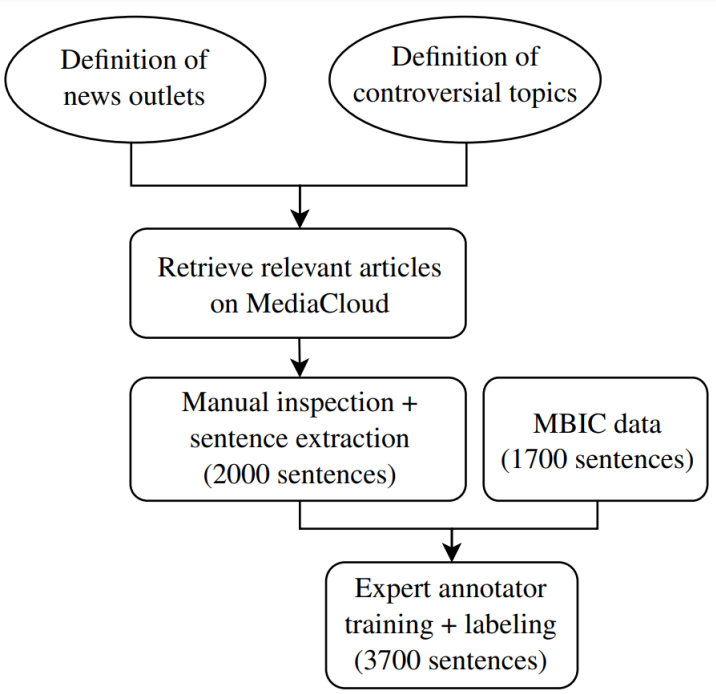

Figure 3.1: Data collection and annotation pipeline of **BABE**, reprinted from [?]

3.3 Wikipedia datasets

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

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

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

When neutrality is contested, Wikipedia article can be moved to NPOV dispute by tagging it with {{NPOV}} or {{POV}}[7] template. Debate on

[6] `https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`
[7] Other POV related variations are ofthen used.

specific details of neutrality violations is then initialized among editors and eventually resolved, leading to removal of the tag.

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

Even though these datasets introduce a large amount 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. Effect of this domain gap on a training of a model is studied in **??** section.

3.3.1 Wiki Neutrality Corpus

Wiki Neutrality Corpus (WNC) [**?**] is a parallel corpus of 180k pairs of biased and unbiased sentences. For the collection of the data, **??** approach was adopted. The authors crawled revisions from 2014 - 2019. Each revision has been processed to check if it contains any variation of *POV* related text in it. 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 which have not been changed during the NPOV dispute were extracted as neutral and for word-level classification purposes, a subset of WNC corpus, where only one word is changed in the biased-unbiased pair, were added.

3.3.2 CW-HARD

Hube et al. [**?**] constructed a dataset based on NPOV, where only revisions with one sentence diff were filtered. However, because of the potentially noisy outcome, 5000 sentences were sampled and annotated using crouwdsourcing. Yet, the Krippendorffs Alpha agreement score measured only $\alpha = 0.124$ which is generally considered low.

After filtering out sentences which annotators labeled with "I dont know" option, the final dataset consists of 1843 statements labeled as biased and 3109 labeled as neutral, a total of 4953 sentences.

3.3.3 WikiBias

This is the latest dataset based on Wikipedia. The authors [**?**] closely follow the approach of WNC [**?**] 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, introduced **WikiBias-Manual** dataset consists of 3400 biased and 4798 neutral sentences annotated with high IAA score of Cohen's $\kappa = 0.734$

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 3.2: Comparison of all bias related datasets collected

3.4 Unused datasets

Some datasets focus on a slightly different task, yet still carry potentially useful information. Such data can be useful in a Multi-Task setting **??**. To name a few, which are focused on a detection of ideology:

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

3.5 Summary

In the previous section, 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 **??**.

BABE dataset is generally a good benchmark and its translated parallel version will be used for evaluation in this work. Combinations of different datasets merging are studied in **??**. Unfortunately a lot of the data suffer from noisy labelling and low IAA greatly. Hence, their usability is considerably limited.

Chapter 4

Czech datasets

Czech language is a so-called **low resource language**, which, in the machine learning community, means that, for a particular language, a limited number of datasets of sufficient quality and size, is available. Thus, the bias detection task in Czech environment is complicated. Despite the relatively sufficient number of datasets in English, there is essentially no Czech one suitable.

In essence, three options to solve this problem are feasible. The most promising way is to annotate a new gold-standard dataset. However, media bias is a nontrivial, 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. **Allsides**[1] for example, 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 [**?**]. Unfortunately, there is no Czech site that would provide **useful** bias information on neither source or article level. Server **Nadační fond nezávislé žurnalistiky** (NFNZ) [2] provides 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, automatic creation of a dataset can be done in a clever way like described in section **??**. Despite the limitation caused by the size of the particular Wikipedia, this approach is suitable for Czech environment, since Czech Wikipedia has a comparably large editor base[3] ranking #26 in a number of edits worldwide. I took this approach and I present a **new parallel corpus** for bias detection based on Czech Wikipedia **??**.

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 **??** using **DeepL**, and finally processed them into a unified format

[1] `https://www.allsides.com/unbiased-balanced-news`
[2] https://www.nfnz.cz/
[3] `https://en.wikipedia.org/wiki/List_of_Wikipedias`

??. .

4.1 Machine Translation

Since 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, as other fields of Natural Language Processing (NLP), has experienced a massive boost in performance, due to the rise of attention mechanism and complex transformer architectures **??**.

Modern machine translation models use the **encoder-decoder** architecture (usually more encoders and decoders stacked on top of each other **??**), where the encoder part distils (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 translation of datasets I chose **DeepL** translator, which is purely[4] NMT based system which outperforms other translation systems by a large margin.

4.2 Processing

For convenience, 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.

Moreover, all sentences, which were originally cased were **not** lower-cased.

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

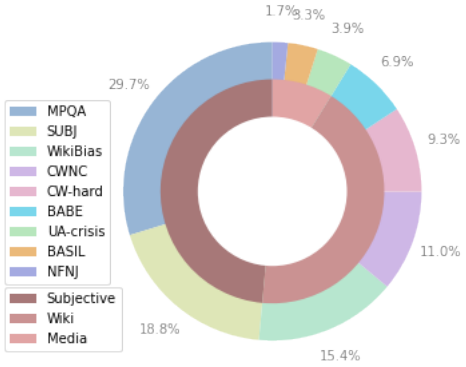

Figure 4.1: Dataset distribution in czech collection of data (Without WNC)

4.3 Translated data

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

I share all listed datasets on HugginFace[5] hub.

- BABE-CS
- Basil-CS
- WikiBias-CS
- CW-hard-CS
- MPQA-CS
- NFNJ-CS
- SUBJ-CS
- UA-crisis-CS
- WNC-large-CS[6]

Together, approximately 400k of bias-labeled translated sentences were collected. A distribution of the datasets can be seen in figure **??**. The WNC is not included in the plot because it represents 87% of all data.

4.4 Czech Wiki Neutrality Corpus

Finally, I present two novel parallel corpora extracted directly from Czech Wikipedia. To the best of my knowledge, this is the only original Czech

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

[6] additional *large* is added for distinction between large translated WNC and czech version of WNC

dataset related to media bias detection. The only partially relevant dataset is **SubLex**[**?**] which is a subjectivity lexicon mainly focused on sentiment. However, lexicon-based approaches are nowadays outperformed by neural models.

I followed two main existing approaches, both of them relying on the extraction of revisions that includes the {{NPOV}} tag or its variation. The NPOV tag has also 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.

4.4.1 CWNC-noisy

I closely followed the [**?**] approach and used their 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 edits containing one of the NPOV related tags are extracted and then the process of sentence extraction follows. This approach yields 15k sentences, however, it uses a rather trivial assumption that when NPOV tag is removed, **all** removed sentences are biased and all added are unbiased. This annotating strategy led to poor results and for this reason, I excluded this dataset from further experiments entirely.

4.4.2 CWNC

This dataset was created following [**?**] approach. The process is the same as described in section **??**. I used **20220201** snapshot of Wikipedia dump, which was, at the time of dataset collection, the latest snapshot that included all 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.[9]
2. All cases has been preserved.
3. Czech Morphodita tokenizer was used[10]

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 **??**

[7] `https://dumps.wikimedia.org/cswiki/`
[8] `https://github.com/rpryzant/neutralizing-bias`
[9] Regular expression used to match npov related comments:
[10] `https://ufal.mff.cuni.cz/morphodita/users-manual`

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 4.1: Example of pairs of biased sentences and their rewritten neutral form

4.5 Not translated

Due to a big size of some data, I was not able to translate more than one large-scale dataset. For this reason NewsB dataset has not been translated, since it is one of the few datasets that doesn't focus on the very same task.

Chapter 5

Theoretical background

In this section, I briefly introduce state-of-the-art methods and models used in experiment section. Many of the problems in NLP are tasks of mapping one sequence to another, therefore modern architectures were designed to tackle this very problem |**efficiently**.

5.1 Text representation

The NLP models operate on 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 set of tokens).

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

After the tokenization, a numerical representation for the tokens has to be obtained. A naive way is to use mapping for every word to an index in predefined/obtained vocabulary. Such an approach suffers from the explicit ordering of the words, which may negatively influence the model. Another possible representation is the **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 representation of words are word **embeddings**. Embeddings are fixed-size feature vectors. The dimension of the embeddings is a hyperparameter, usually with the value between 100 and 1000. Embeddings are usually learned along the particular task, in an Embedding layer. But there are also precomputed representations available, such as word2vec, Glove or ELmo.

5.2 Neural Networks

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

$$f(x) = \phi(w^T x) \tag{5.1}$$

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 a 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.

5.3 Encoder-Decoder

Vanilla NN architecture operates on inputs of fixed length. For variable length (for example a sentence) input 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 a **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.

Sequence-to-sequence [**?**, **?**] 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 to 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 sequence is "forgotten". Secondly, RNNs work in a sequential manner, hence there is no room for parallelism. A **Transformer** architecture aim to solve these problems.

5.4 Transformers

Transformers is a family of deep neural architectures, that has revoulutionized an NLP field. The **attention** mechanism is a core principle behind inside transformer architecture and solves some of the problems mentioned in previous section.

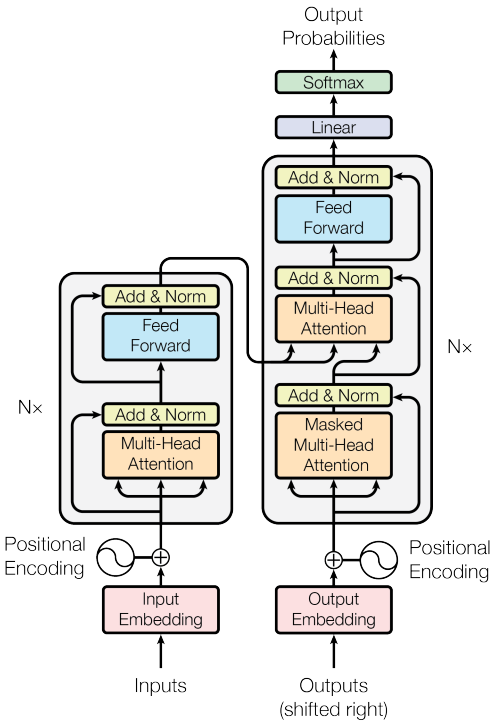

Figure 5.1: Transformer model architecture, preprinted reprinted from [**?**]

Current State-Of-The-Art (SOTA) language models, such as GPT-3, BERT, RoBERTa all use this architecture with different modifications (eg., BERT only uses the encoder side etc). For detection of media bias, I have eventually also narrowed the choice of models to transformers rather than other neural architectures.

5.4.1 Attention

Attention [**?**, **?**] is essentially a mechanism that alows the unit (either it is on the decoder or the encoder side) to learn to focus on some segments of input sequence more than the others. This mechanism has been motivated by the problem of "forgetting" of information in a long sequences mentioned before. This way, parts of the sequences that potentially drives the decision are represented more in hidden states than other parts.

5.4.2 Transformer architecture

Scheme of the original architecture can be seen in **??**. Just as previous seq2seq models, transformer also consists of Encoder and Decoder modules. Precisely, each module is a stack of n identical encoder or decoder layers.

Each encoder layer

Instead of working on each token sequentially (as conventional Encoder-Decoder did), transformer builds hidden state representations for all tokens in sequence in **parallel**.

5.5 Text classification

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

$$L_{BCE} = \frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot (1 - \hat{y}_i)) \tag{5.2}$$

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

5.5.1 Metrics

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

$$accuracy = \frac{correct\ predictions}{total\ predictions} \tag{5.3}$$

This metric is feasible if classes of the dataset are balanced. However, imagine a situation where 90% of data belong to one class and only 10% to another. Classifier, that 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*.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.4}$$

where precision

$$^{1}Precision = \frac{TP}{TP + FP} \tag{5.5}$$

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

$$Rrecall = \frac{TP}{TP + FN} \tag{5.6}$$

can be understood as "how much of a positive class can model predict". Scores for each class are then averaged to obtain the final score.

5.5.2 Transformers for text classification

Predictive power of transformers is behind every SOTA result and text classification is no exception. For classification, usually only the Encoder part of a transformer is used, although some define the classification problem as a sequence-to-sequence and incorporate the decoder too [**?**].

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

During tokenization, 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 the difference, that when the forward pass reaches the classification layer, only [CLS] token is passed as an input. [CLS] token can therefore be understood as sort of a sentence embedding.

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

5.6 Transfer learning

Nowadays the true power of transformers lies in **transfer learning**. Transfer is a process when some knowledge is not learned from scratch but transfered from previously trained model. Since large language models such as BERT or RoBERTa has millions of parameters, it would be extremely costly to train them from scratch.

Such large models are usually pretrained on a very large 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, where random sample of tokens in input sequence is replaced with [MASK] token and model learns to predict the original token.

Having such pre-trained model, one can then easily train a task specific head on top of pre-trained representations. Although process of **fine-tuning** where all parameter of model are tuned, is more often adopted.

5.7 Multi-Task learning

Chapter 6

Experiments

In this section I present experiments on text classification over collected datasets. The main **target** dataset for evaluations is BABE, due to its high quality and properties.

Due to the novelty of CWNC I also perform a baseline evaluation on this dataset, but furthermore it is not tuned. I follow the current standard approaches and use pretrained transformers for further pretraining and fine-tuning.

A brief summary of the models tested can be found in the following:

6.0.1 Czech monolingual models

- **RobeCzech** [**?**] - RoBERTa-based model with 125M parameters. Just as its original counterpart, it is trained with MLM (Masked Language Modeling) task, on 4,917M tokens of czech corpora.
- **Czert** [**?**] - BERT-based model with 110M parameters, trained with MLM and NSP (Next Sentence Prediction) tasks. All Together trained on 37GB of text.
- **FERNET-C5** [**?**] - BERT-based model trained with the MLM and NSP task on 93GB of text from the Common Crawl project.
- **FERNET-News** [**?**] - RoBERTa-based model trained with MLM task on 20GB of Czech News text.

6.0.2 Multi-lingual models

- **SlavicBert** [**?**] - 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** - BERT-based model with 179M parameters trained on corpora of 104 languages, Czech included, with MLM task.

target\models	Czert	RobeCzech	mBERT	FERNET-C5	FERNET-News	SlavicBERT
BABE	0.776	0.774	0.734	**0.781**	0.566	0.754
CWNC	0.735	**0.765**	0.730	0.719	0.593	0.732

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

6.1 On Instability of fine-tuning

bla bla. Cross validace , ukazalo 10x menší variance. -> 10fold cross validace In turned out that different initializations perform differently on different splits. Therefore increasing number of folds in cross validation reduced the variance 10 times.

6.2 Experimental setup

All models are fetched, trained, and evaluated using the HuggingFace API. The maximum sequence length is set to 128 tokens. All the parameters can be seen in the Appendix. Everything is evaluated using 5-fold cross-validation with fixed seed. The evaluation metric for all experiments is F1 score with macro averaging. A small portion (15%) of the target dataset is left aside as a **test set** at the beginning and used only for the final evaluation to ensure that there are no data leaks.

First, the baseline models are evaluated, and the best model is selected for further compact hyperparameter tuning. Then experiments on different combinations of datasets are then performed. All training has been done on a single GPU on RCI cluster.

6.3 Baseline setup

As a baseline, all Czech models are fine-tuned on BABE and evaluated using 5-fold stratified cross-validation. Hyperparameters were the same as those used by the authors [**?**]. However, the authors used early stopping together with cross-validation and used the validation split inside CV to early stop, which is not ideal since the split should be used only for evaluation. This way the model can "see" the data before evaluation; hence I did not use early stopping with CV at all and fixed the number of epochs to 3, as authors of BERT suggest [**?**] . All other hyperaparemeters remained unchanged. AdamW optimizer is used with an initial learning rate 5e-5.

Baseline evaluation of all the Czech models used can be seen in table **??**. The final F1 score is averaged across all folds. For further experiments and tuning, I chose the model that performed best on average between the two datasets, which is **RobeCzech**.

6.4 Hyperparameter tuning

I restricted the search space only to the combination 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. After running the grid search, the overall best parameters were:

$$\{\text{learning\_rate} = \text{2e-5,batch\_size} = 16, \text{ epochs=3}\}$$

6.5 Combining Datasets

This section is dedicated to the study of the influence of pre-training and training with different combination of datasets. Trying all combinations would result in training of 128 models, which is obviously infeasible. Therefore, few arbitrary combinations of the datasets with respect to their bias information were sampled and evaluated.

6.5.1 Trained on Datasets, evaluated on BABE

As a starting point, I trained a model on each of the datasets and evaluated their performance on BABE and CWNC, to see, which datasets provide the most relevant bias annotations with respect to the targets. Data from wikipedia family perform comparably better, mainly because their bias information is more straightforward. On the other hand, low quality of MB datasets proves to be problematic. In most cases, pretraining on small dataset only hurts the performance. Only in case of WNC which is large-scale dataset, pretraining outperformed baseline of CWNC. That is because both data come from the same distribution.

6.5.2 Pretrianing Combinations

- **All datasets** - I pretrained all to see which are generally best. Trained all and evaluated on babe, to see what is the most relevant. On the same valid splits. Low quality of media bias data really doesnt work out. Too small to learn bias embeddings. Pretrained odchylka is too little, it has no impact at all. Wiki works very good. On average variance between all tasks was so low that they bbasically performed tall the same. Thus pretaining on these smaller datasets made no improvement.
- **Subjective Bias** - pretraining had no effect. Subj achieved good performance. Subj did actually quite good.
- **Media Bias**

	baseline	SUBJ	WIKI	MB	WNC	ALL
Pretraining + Finetuning	0.7835	0.7875	0.7797	0.7702	0.7825	**0.7878**
Pretraining + Evaluating	-	0.5542	0.6344	0.4631	**0.6697**	0.6423

Table 6.2: F1 scores of baseline fine-tuning. Best scores for each dataset are highlighted.

- **Wiki Bias**
- **WNC** - Since this is the only large dataset I decided to evaluate it seperately.
- **All together** - To learn bias embeddings properly.

conclusion: pretraining had no effect,Finetuning on small datasets is very unstable weight inicializations are bad.

Training wiki, freezing parameters nad only training classification layer. Freezing encoder havent worked out yet!

??

6.6 Final training

6.7 Self-Training

WikiBias uses selftraining. I use selftraining, stonks. I use intuition behind early stopping procedure and extrapolate it to the training. The model is contually I experimented with few datasets. Self-Training is usually performed on arbitrary corpora of text. I decided to incorporate the translated datasets, because they were manually selected as a representative sample of the class, I believe that using regular news corpora would include too many neutral examples, thus would be less efficient.

6.8 Notes on experiments

For better evaluation nested cross-validation should be used. However, the computational demand would be too large. Random seed was not enough. PyTorch backend brought in some inner randomness. A full study with more models could be performed, but that would require enormous number of trained models thus i perform all experiments on RobeCzech.

6.9 Inference on Czech News Samples

- Analysis and statistics
- Few words Article level

Chapter 7

Conclusion

7.1 Summary of work done

In this work I collected and analyzed all the literature and resource for studying state-of-the-art media bias detection. I present new czech parallel corpus derived from wikipedia and in addition 9 parallel translated czech datasets for tackling the media bias detection in Czech language.

I trained and tuned the state-of-the-art language models to achiave F1% score of ... on target dataset, which is comparable to current SOTA in English.

Finally, the final classifier has been used to build a demo and to analyze a sample of articles from cesky rozhlas throughout the history. Results of this study showed interesting information about progression of media slant in Czech news.

I hope my work will kickstart the research of media bias in Czech news and will motivate future researchers to build up on this work and potentially build larger and better corpora.

7.2 What hasnt make it into this work

Even though MTL approach seems promising, it would require a lot of work on task selection and task evaluation, which is not focus of this thesis.analysis of decisions, creation of ground-truth dataset.

7.3 Future perspective

My experiments suggest that self-training can improve performance on low resource language with small size dataset. There are more sophisticated methods of sampling which could be used.

As discussed in the experiments section, reasearch suggests that multitask learning increases classification accuracy significantly ref. Multitask model environment requires a lot of tasks [**?**] to perform better than single task models. Therefore, for czech language setting, one of the future research possibilities would be to leverage multi-task learning for current classifier improvement.