

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

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.
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

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 L^AT_EX class, what are the requirements, etc.

Keywords: bias detection, transformers, text classification

Supervisor: Ing. Jan Drchal, Ph.D

?abstractname?

Tento manuál představuje L^AT_EXovou třídu ctuthesis, její použití, požadavky na systém atd.

Keywords: bias detekce, transformers, text classification

Contents



Figures

Tables





Chapter 1

Introduction

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



1.1 Motivation

tady něco o nlp



1.2 Outline

Chapter 2

State of the art

2.1 Bias

Mitigation vs Detection, importance, types of bias

2.2 Gender bias

mdgender = classification. stereoset (and others) = mitigating? Experiments in experiments section

2.2.1 md_gender dataset

2.3 Media bias detection

2.3.1 Informational vs Lexical

Many do lexical, framing, něco o dalších, WCL.

2.3.2 Methodology (SOTA)

bla bla Article level vs sentence level. Neural nets vs classical machine learning. Multitask learning.

Chapter 3

Datasets

Due to the varying definition of bias, many datasets aim to detect different angles of bias. In this section, I present a collection of all datasets available, related to biased writing and subjectivity detection.

Because there are not many media bias datasets of sufficient quality, I decided to gather all relevant data which are on some level related to the media bias and later leverage their bias information to augment smaller ground truth datasets, which I will discuss in the experiment chapter.

As discussed in ?? this work only focuses on sentence level classification, thus datasets on the article level are not considered.

■ 3.0.1 SUBJ

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, hence the labels can be assumed to be noisy. The authors made an assumption that all reviews from www.rottentomatoes.com are subjective and all plot summaries from www.imdb.com are objective. Then 5k of sentences were sampled randomly for each class.

■ 3.0.2 MPQA

Multi-**P**erspective **Q**uestion **A**nswering (MPQA) Opinion corpus is another dataset that can be used for subjectivity detection. For the purpose of our task, I used the MPQA Opinion corpus version 2.0, which consists of 692 articles from 187 different news sources summing up to 15,802 sentences. All articles are from June 2001 to May 2002. (+ topics?).

The corpus offers a rich annotation scheme [?] that focuses on sentiment and subjectivity annotations. For bias corpus creation, I focused on two types of annotations:

- Direct subjective
- Expressive subjective

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/biased. All other sentences were extracted as objective/unbiased.

This approach yielded 9,484 subjective sentences and 6318 objective sentences.

■ 3.0.3 BASIL

BASIL dataset [?] consists of 300 articles with 1,727 sentence level bias annotations. The authors of the dataset distinguish between **lexical** and **informational** bias. They define lexical bias as a form of bias which does not depend on context and usually introduce polarized words.

The annotations are performed by two experts and further resolution discussions later lead to 0.56 agreement score for lexical bias.

Even though BASIL dataset brings the sufficient annotation quality, most of the labelling resulted in informational bias annotations, leaving only 478 sentences with lexical bias information. Therefore, the data are usable only for evaluation.

■ 3.0.4 Ukraine Crisis Dataset

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

The authors offer rich annotations for each sentence. Each one of them looking 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 totalling of 44 547 fine-grained annotations. For media bias detection in the experiment chapter, only overall bias annotations were used. Even though this is one of the highest quality dataset regarding media bias specifically, it also suffers from low Krippendorff's alpha score (-0.05). Hence, its usability is limited.

■ 3.0.5 NFNJ

The NFNJ¹ dataset provides 966 sentences from 46 articles with annotations on a fine-grained level. Despite the relatively small size of the dataset, the

¹[?] refer to this dataset as NFNJ, however in the original paper the name is not presented.

Inter-Annotator Agreement (IAA) measures Fleiss Kappa scores of zero on average.

Authors share the dataset for research purposes, however, the public version differs from the one described in the original paper. For creating the final data I made a few assumptions:

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

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

To obtain final truth labels in neutral/biased format I simply assumed sentences with score ≤ 2 as neutral and > 2 as biased.

■ 3.0.6 BABE

A key media bias dataset from Media Bias Group (MBG), which is to my best knowledge and according to the authors, 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 for 2000 more sentences, altogether covering 12 topics.

BABE has been annotated by 8 experts resulting in IAA Krippendorfs $\alpha = 0.46$, which outperforms other media bias datasets by a large margin. The dataset also provides detailed information about annotators background making it a reliable source of bias information. The scheme of collection of sentences and labelling can be seen in ??

■ 3.1 Wikipedia NPOV datasets

Due to annotation costs and the overall lack of large-scale datasets in media bias settings, many researches [?, ?, ?] used Wikipedia's Neutral Point Of View (NPOV) policy². to construct large-scale datasets automatically.

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

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

²https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

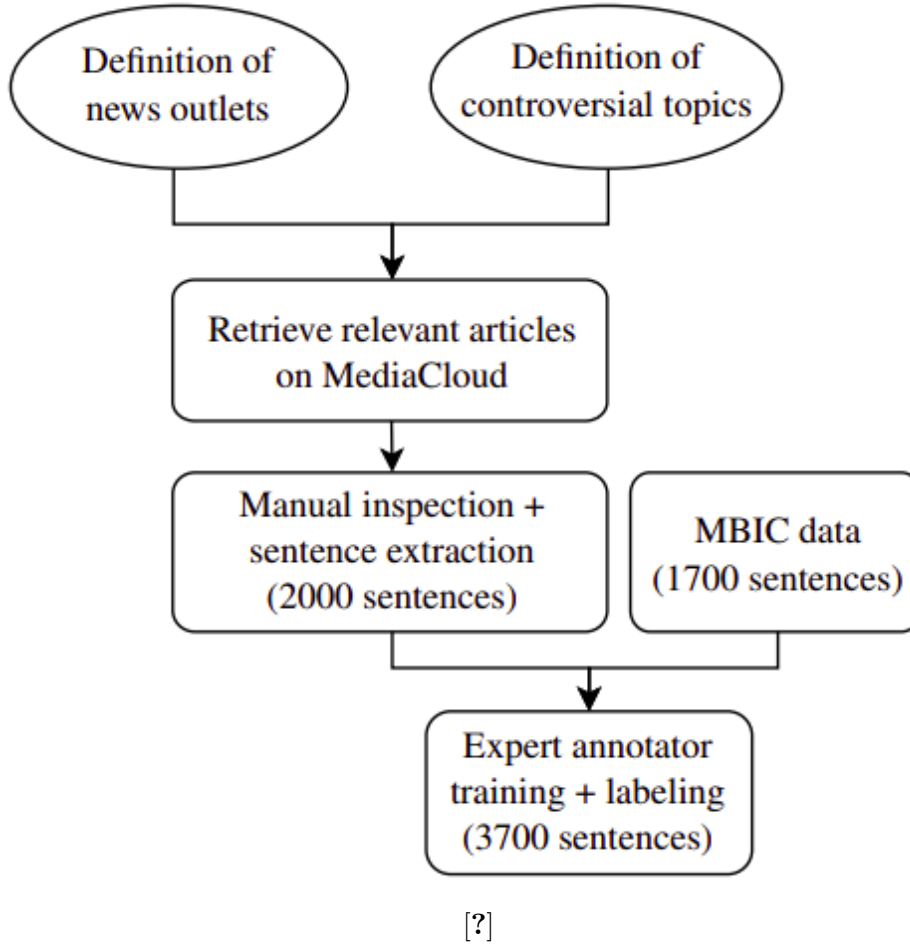


Figure 3.1: Data collection and annotation pipeline

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

This bias information can be used to extract parts of text, that violates 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 points were unbiased.

Even though these datasets introduce large amount of samples that are highly related to media bias, they are all sampled from Wikipedia’s environment, which isn’t same as the News environment.

■ 3.1.1 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 of wikipedia from time span 2014 -

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	3700	annotators	medium
WNC	362.990	automatic	-
CW-hard	4953	crowdsourcing	low
WikiBias	8198	annotators	high

Table 3.1: Comparison of all bias related datasets collected

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 sentence before edit is considered biased and modified/added sentence after edit is considered to be neutral/unbiased.

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

■ 3.1.2 CW-HARD

Hube et. al [?] constructed a dataset based on NPOV, where only revisions with one sentence diff were filtered. However, this leads to a very noisy dataset, thus the authors sampled 5000 sentences and used crowdsourcing to annotate them with bias/unbiased labels. However, the Krippendorffs Alpha agreement score measured only $\alpha = 0.124$ which is considered low.

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

■ 3.1.3 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 higher quality corpus, 4,099 sentence pairs were randomly sampled and labeled by trained annotators. As a result introduced WikiBias-Manual dataset consists of 3,400 biased and 4,798 neutral sentences annotated with high IAA Cohen’s $\kappa = 0.734$

■ 3.2 Unused datasets

Several datasets that are related to media bias but not suitable for the media bias detection task.

■ 3.2.1 NewsB

Focused on detecting political party/ideology.

3.2.2 IBC

Is a dataset focused on detectino of ideology. However,it is not publicly available and I was not able to get the dataset from the authors.





3.3 Datasets summary

In this section I introduced all datasets that are publicly available and are more or less related to the media bias detection task. The overview of all datasets and it’s properties can be seen in figure ??.



Chapter 4

Theoretical background

-  4.1 Text classification
-  4.2 Neural Networks
-  4.3 Attention and transformers
-  4.4 Transfer learning

Chapter 5

Czech datasets

Despite the relatively satisfying number of datasets, there is essentially no Czech dataset which is at

5.1 Translation

5.1.1 DeepL

5.2 Processing

Since BERT can handle cased.

5.3 Analysis

output unified set of Czech bias related datasets **Czech Unified set of Bias Data**

1. mpqa-cs
2. subj-cs
3. newsb-cs
4. cw-hard-cs
5. wiki-npov-large-cs
6. babe-cs

5.4 Czech Wiki Neutrality Corpus

Finally, I present two novel parallel corpora extracted directly from Czech Wikipedia. To my best knowledge, this is the only original Czech dataset related to media bias and subjectivity detection. The only related dataset is SubLex which is a subjectivity lexicon mainly focused on sentiment. However, lexicon based approaches proved to be insufficient in tackling complex media bias.

For the dataset creation, 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.

WIKI1-CS

For this dataset I followed the [?] approach and their script. First, a file with all pages and its complete edit history is downloaded from wiki dump. 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. All used tags can be seen in appendix.

This approach yielded 15k sentences, However, it uses rather trivial assumption that when NPOV tag is removed, all removed sentences are biased and all added are expected to be unbiased. This annotating strategy later [odkaz na experimenty] proved to be insufficient and yielded very noisy dataset. For this reason I excluded this dataset from further experiments.

WIKI2-CS

This dataset was created following [?] approach. The process is the same as described in section ???. I used "20220201" snapshot of wikipedia dump. I chose the latest version that included all the necessary files. I used the script publicly available on github [odkaz na repo], with few slight modifications to fit the czech language properties:

1. Regex was extended to exclude czech words that contain "pov" inside eg. povstání, povlak etc. ¹
2. All cases has been preserved, since bert like models can handle cased language.
3. Czech Morphodita tokenizer was used.

¹Regular expression used to match npov related comments:

Before: od roku 1993 byl dvě funkční období prezidentem české republiky, kterou vyvedl z varšavské smlouvy a **navzdory svým slovům** dovedl do nato.
 After: od roku 1993 byl dvě funkční období prezidentem české republiky, kterou vyvedl z varšavské smlouvy a dovedl do nato.

Figure 5.1: Example of CWNC sentence pair

Final dataset consists of:

1. 3k of "before" and "after" sentence pairs
2. 1.7k subset of mentioned set where only one word was changed
3. 7.5 sentences, where the change was rejected or reversed implying neutrality of the original sentence

The random example of sentence pair can be seen in ??

5.5 Not translated

Since my current DeepL plan allowed me to translate only one "large-scale" dataset due to the DeepL's fair usage policy I decided to not translate the NewsB since it focuses on distinguishing between conservative and liberal bias hence is not directly applicable on our task. However the hyperpartisan task is a good candidate for multi task setting as suggested later in ?? section

Chapter 6

Experiments

6.1 Czech models

RobeCzech

1. RobeCZECH
2. Czert
3. FERNET

zmínka RCI cluster

1. Training args tuning
2. Weights and biases visualisations
3. Pre-finetuning, self-training
4. experiment on collective dataset

6.2 Evaluation


6.3 Inference on Czech News Samples

6.3.1 Analysis and statistics

6.3.2 Few words Article level

6.4 LIME analysis and demo

6.5 Multi-Task learning approach



Chapter 7

Conclusion



7.1 Summary of work done

In this work I presented 8 parallel czech datasets for tackling the media bias detection. 6 of which are related to subjectivity detection and one is.



7.2 Future perspective

As discussed in the section [experimenty] reasearch [citace] suggests that multitask learning increases classification accuracy significantly ref. However, multitask model environment requires a lot of tasks [odkaz na exT5] to perform better than single task models.