Dr hab. inż. Maciej Piasecki
Katedra Sztucznej Inteligencji
Politechnika Wrocławska
maciej.piasecki@pwr.edu.pl

20th November 2023

Reviewer's opinion
on Ph.D. dissertation authored by
Katarzyna Baraniak
entitled:
Machine Learning Tools and Techniques Supporting News Media Bias Analysis

1. Problem and its impact

What is, in your opinion, the most important problem discussed in the dissertation?

The main goal of the thesis has been declared "to propose methods that can help in news media bias analysis". This plural number occurring in the subject determines the thesis character. Its main theme is news media bias analysis, but not the main problem in a strict meaning. The thesis discusses several different methods for specific problems related to the theme, concerning its different aspects, but, finally, the methods are not presented as comprising a complete approach. The different subproblems are aligned along the identified three main dimensions of the news bias, but they are treated independently, to such an extent that they are not tested, both alone and in combination, on the same datasets, i.e. even in the conclusions it is not shown how to analyse one single news set with the proposed bunch of methods and what could we learn from their combined application. It is true that the issue of news media bias analysis is quite broad, but it seems that the logic of the thesis' organisation has been dictated by the series of research papers published by the author, presenting the different techniques, and related to the subsequent chapters.

Thus, we should rather focus on the particular subproblems discussed in the thesis. Two of them are the most prominent: entity-level sentiment analysis in short political news texts and persuasion techniques detection.

The problem of entity-level sentiment analysis is in fact a specific narrowly treated subproblem of aspect-based sentiment analysis. This has not been clearly enough explained in the thesis. For instance, this relation is almost neglected in the comparison to the literature. This subproblem is presented as "the entity-level sentiment of news headlines from readers perspective", but the analysis from the readers perspective is standard as it is the only one possible when we take into account that annotation of the training-testing datasets is always done by readers, not speakers of the utterances.

For the second prominent subproblem, namely persuasion techniques detection, an interesting approach based on hierarchical neural networks was proposed. However, it seems to be separated from and to not be connected to the main theme of news bias analysis, unless some far going

further interpretation has been done by the reader. The presented discussion on the definition of news bias is quite brief. The assumed way of defining it in terms of three aspects: visibility, tonality, and agenda, is a good point of departure for a computational approach, but still the persuasion techniques detection is not explicitly related to any of the three aspects, and this exercise is left for the readers' intuition.

In addition to the main subproblems, the thesis presents several "helper techniques", like news articles similarity detection, entity timeline analysis and news source detection. However, concerning the methods, they are relatively simple and mostly applications, while their combination into a more complex analysis does not go beyond a kind of proof of concept and, unfortunately, does not delve deeper into the interesting research area of the analysis of temporal text streams. It is a pity, because this could be a very good framework for the whole thesis and the subproblems.

Is it a scientific one?
The main theme and all the subproblems addressed by the author have scientific character. Even in the case of the "helper techniques" they are applied and discussed from the perspective of application in research.

The way of approaching the main theme and subproblems is also scientific in nature, even if being quite shallow in several cases and aspects.

Does it have a practical meaning?
The main theme, as well as the two subproblems in focus may have very practical meaning, e.g. as research tools or tools for media analysis. The presented work does not go so far, but all software has been released on open licence – very good move from the point of view of research reproducibility. Moreover, a valuable benchmark dataset – "a novel dataset called SEN for entity-level sentiment analysis" has been introduced. It is important that it is a new dataset, not a compilation of the existing ones.

2. Contribution

What is the main, original contribution of the dissertation?
The contribution of the thesis is related to the three main aspects of news bias, namely: visibility, tonality, and agenda.

Concerning the visibility, simple direct means for, so called, named entity recognition (NER, i.e. recognition and semantic classification of proper name mentions in text) were used but with an interesting application in focus, i.e. analysis of temporal news streams. Unfortunately, because of not going deeper than calculation of direct frequencies, lack of normalisation, only binary model, and only intuitive analysis of the PN frequency charts, the originality of the contribution is limited to the general idea. It is followed by a good illustrative application of combining entity timeline with changes in sentiment.

News similarity detection is proposed as a technique in a similar context. However, relatively basic techniques for finding similar news articles were applied. Training-testing dataset annotation is not specified. Finally, the claimed finding sources for published news articles is definitely too

far going. Thus, here the contribution is very unclear. This part of the thesis includes a set of experiments on application of known methods.

The main thesis contribution lies in entity-based news sentiment analysis and persuasion techniques detection.

Concerning the former it is clearly related to tonality. A well defined and built dataset was introduced, called "SEN" ("Sentiment concerning Entity in News headlines"). It is a valuable contribution to the domain and is based on several different annotation procedures that makes it more reliable. Its value is supported by the data curation done with simple, but effective techniques, and various analysis of the dataset. Its only limitation is fact that the dataset includes only mentions of a small number, preselected PNs of high frequency that results in some bias of the dataset. SEN appeared to be quite challenging dataset for a state of the art an aspect-based sentiment analysis method. Thus, proposed SEN shed some new light on this well studied problem.

A wide range of methods was tested for the entity-based news sentiment analysis, among them a novel method called EntBERT and methods based on utilisation of several layers of a transformer. EntBERT is a simple, but promising modification of BERT-based classification "a multitask hierarchical BERT-based neural network". It showed good performance on the proposed, difficult dataset.

The author observed and studied an issue of entity bias in an insightful set of well documented experiments. Several variants of transformer-based methods utilising proper name representation enhanced with an external textual knowledge source were proposed and studied.

A novel work approach to persuasion techniques detection was proposed. It is based on an interesting application of multitask learning combining two tasks of different character: span identification and persuasion techniques identification – sequence tagging and the latter short text multilabel classification. They have been combined within a hierarchical neural network whose two components are associated by the loss function and representation passing. In spite of being based on known elements and tested only against a transformer-based baseline, the method is a valuable contribution.

3. Correctness

Can we trust what is claimed in the dissertation?
The dissertation is written in a systematic and careful way. All the decisions and steps are generally well motivated and described. However, trust to the findings should be unconditional due to the limited comparison to the state of the art (in many places) and also not deep enough experimental evaluation (in several cases).

In the case of tracing proper name mentions, the problem of limited accuracy of the NER methods was neglected. No discussion of the influence of identification error or its correction was included. There is also no attempt to perform entity linking to some knowledge resource, in order to solve potential ambiguities. In the case of the collected frequency data, there is lack of normalisation, other than the applied binary model – but still used implicitly. Only intuitive

analysis of the proper name frequency charts is given with no mathematical statistics methods applied. It is a pity that methods for comparing corpora were not applied, too.

For studying news similarity a different dataset was used and this is reoccurring problem, that almost every part of the thesis is referring to a different dataset. In this case training-testing dataset annotation is only mentioned but not specified. All three experiments present a rather pre-matured level, and are not reliable, more illustrative examples.

SEN ("Sentiment concerning Entity in News headlines") seems to be a valuable resource. The only problem is mixing two different issues in the annotation guidelines: opinion and sentiment, e.g. (p 61) "sentiment may be revealed by clear statement or opinion about an entity".

In the case of sentiment analysis, methods based on the proper name positions, i.e. Target-Dependent LSTM (TDLSTM) and Target-Dependent BERT (TDBERT) may be affected by implicit bias of explicitly marking the proper name position – there is no perfect NER to do this automatically. So, the real performance in practical tasks may be significantly lower. Moreover, representation of a target by its subtoken vector is often used in relation recognition that has not been mentioned in the thesis.

In the case of EntBERT one small unexplained issue, or rather a curiosity, are permanently worse results of EntBERT expressed on SEN-en including headlines in English.

The proposed utilisation of representations from all layers of the transformer is a good move. Many studies in literatures revealed diversified properties of BERT-like model layers, but such works are not mentioned in the thesis. However, it is intriguing why it is only LSTM that has been applied as a classifier? The vector of CLS subtoken representations from all layers is always of a fixed size and is not of sequential character, at least in the same sense like a sequence of words. Other types of classifiers, e.g. CNN, MLP, have not been tested. Limited scope of experiments (with the exception of sentiment recognition in Chapter 5) and lack of ablation studies for the proposed solutions are general problems of the thesis.

Aspect-based sentiment recognition has been intensively studied for many years in literature, entity-based sentiment recognition is only its subfield, but comparison of the proposed solution to the state-of-the-art methods is very limited in the thesis. This is even more striking in the case of persuasion techniques detection (Chapter 6) where there is no comparison at all. The only comparison is made with a baseline and a typical BERT-based classifier.

The sentiment annotated datasets have been divided into the train and test part, but nothing was said about the development parts used for tuning the model and training process parameters.

In the case of enhancement with "external context", there is an implicit assumption that an entity mention in text can be unambiguously linked to the appropriate description in a knowledge resource, in general this is not true. Neither the influence of inevitable ambiguities, nor entity linking techniques for solving it are discussed.

Comparison of the models utilising information from all layers of BERT transformers in Tables 5.10 and Table 5.11 is difficult to follow and not fully convincing, as no baseline is shown in both tables and only the unmasked (i.e. biased) version of the dataset is used. Moreover, in all

experiments using the unmasked dataset there are no attempts to guarantee lexical split property, i.e. to have different set of PNs as targets in the training and testing part.

EntSeqBERT2 is described as building representations from sequences of "entities in the whole article", while it is reported to be applied to SEN-pl and SEN-en in which only single headlines, not whole articles are included.

In addition, the thesis concentrates on headlines responsible for the first impression of the reader, but news content is also very important for bias analysis. It is not clear why it has been omitted in the thesis.

During the last couple of years, large generative language models (LLM) showed their ability to analyse text properties, but there is no comparison in the thesis to any technique based on a generative LLM, e.g. by applying so called prompt-based learning, e.g. at least with ChatGPT.

In Chapter 6, in the case of persuasion techniques recognition, there is no clear comparison to a non-multitask learning setting, and comparison to any other method is almost neglected, as it was already mentioned. Ablation studies are missing, e.g. in relation to annotation schemes, representation, loss function etc. Such lacking elements decrease trust in the presented results.

The thesis as a whole fails to address the problem of bias detection (or analysis) in a complete form. It presents several methods focused on selected aspects of the general problem. In addition, the proposed techniques work on different levels of text granularity: articles and headlines. In different methods different datasets are used that makes harder to see effects of their potential joint application.

Are the argument s correct? Indicate the flaws you have noticed, if any.
Generally, narration and thesis structure are convincing, but several flaws can be noticed.

Why was not topic analysis (topic modelling) considered in relation to the agenda aspect of bias detection?

Why tonality aspect of bias has been narrowed down only to the analysis of headlines, very specific type of text?

In Chapter 4, there is no clear distinction between named entity and its mention (or reference by its name) in texts.

Confidence intervals are marked in tables, but there is no precise information on how they were calculated and there is no careful discussion on statistical significance of the observed differences: "We can observe that our model EntBERT outperforms other models on the 3 datasets SEN-en-AMT, SEN-en and PTB.", while the results are very close.

In Chapter 6, how is the problem of persuasion recognition related to the media bias analysis - not clearly explained at the beginning.

IO annotation is the simplest extreme in BIO family of annotation schemes; other annotation schemes were not considered.

4. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of ***Information and Communication Technology****. What areas of that discipline are covered by those chapters/sections? What do you think about quality of those chapters/sections?*

The author showed her good and broad knowledge in Information and Communication Technology, that is especially visible in diversified methods used in solving the problems.

What is your opinion on the list of references? What is the degree of its completeness?

On the general level the list of references is a good background for the discussed issues. Only aspect-based sentiment analysis is treated in very short and selective way.

In a similar way, selection of classification and representation methods in Chapter 4 is not very consistent, and also the most contemporary techniques have been omitted.

5. Other remarks

Additional detailed comments:

p 14: "thearticle"

p 16: "brazilian portuguese"

p 19: zbędne rozważania o bag of words, często to też kolekcja, a nie zbiór

p 20: wrong formula given as idf

p 25 ReLU function is a linear function, but only cut to the positive domain; once again very basic issues are unnecessarily discused

p 28: claiming that vector-based semantic representation appeared only thanks to neural network is too far going simplification; the „word embedding" term was introduced by Mikolov, but dense vector representations, not mentioning distributional semantics, have much longer history

p 43: there are no named entities in text, they are extralinguitic entitites.

p 45: why not to use a text window for the analysis of coincidence

p 51: dlaczego podobieństwo do źródła ma świadczyć o obciążeniu?

p 52: Morfeusz to analizator, daje wszystkie formy, nie ujednoznacznia

p 49: "they encounter in the same article" - grammar?

p 52: "we collect all articles from the specified web news portals. Each article is assigned to a group with articles about similar event." — not always one article is about one event, what in cases this is not so?

p 52: "to the base form using Morfeusz li- brary" — it does not perform lemmatisation!!

p 53: Doc2Vec is based on word2vec, not a very advanced method

p 53: "there are the least false positives." - gram.?

p 53: 4.2.5.1 Group approach — why have not been typical clustering quality measures applied??

p 54: "where the reason may be that it was not able to detect dependencies in long text" — of course, if the lengths were beyond the input window to BERT or other network

p 55: 4.2.6 Experimental Results on News Article Source De- tection — how were the training-testing data and the whole problem defined?

p 57: "entity-level sentiment detection in news headlines" — why only in the headlines, not in the texts if we are looking for bias in media

p 61: "be revealed by clear statement or opinion" — positive or negative stance is something completely different than sentiment; was the stance separated from sentiment?

p 63:++"Furthermore, we analysed the annotators in terms of their bias towards the annotated entities." — very interesting idea

— but $Sentiment\_score$ is simply normalised difference between pos and neg annotations, not a real test for some significant bias towards a particular entity

p 64: Table 5.3 calculating any statistical measure from a handful of annotations is not very scientific!

p 64: "namely tonality bias detection in headlines" — how is this related to sentiment?

p 67: "there are do not exist" — gram.

p 67: LLM is rather used for generative models of complexity much higher than BERT; calling BERT an LLM is misleading

p 69: Chapter 5.3: all results obtained for the target oriented sentiment analysis are very low with LSTM — problems with the task?

p 71: Table 5.7 statistical significance not verified!

p 72: ++nice experiment on testing entity bias

p 74: "at hte index" – a typo

p 76: "a sequence of entities"" – why is plural number used in the case there is always only one target entity? Unclear

p 76: "two LSTM layers that learn about entities." – quite informal and definitely unclear statement

p 92: "If no span is identified the layer takes the first token of BERT embedding." – a mental short cut, BERT embedding does include any token, it is generated for a token sequence, what sequence is referred to here?

p 93: "on dev set according to the micro-f1 metric" — why is micro-F1 used in the situation in which there is significant imbalance of classes?

Bibliography:

Some incomplete records, even describing the own works of the thesis author. Errors in capitalisation of letters in the records.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the *Act of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

☐	☒	☐	☐	☐
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of automated ontology development?

☐	☒	☐	☐	☐
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

☐	☒	☐	☐	☐
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

I hereby recommend acceptance of the thesis of mgr Katarzyna Baraniak for the further steps of the PhD procedures and public defence.

Maciej Piasecki

Signature

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276