

News articles similarity for automatic media bias detection in Polish news portals

Katarzyna Baraniak

Polish-Japanese Academy of Information Technology
Warsaw, Poland

Email: katarzyna.baraniak1@pjwstk.edu.pl

Marcin Sydow

Institute of Computer Science, Polish Academy of Sciences
and Polish-Japanese Academy of Information Technology

Warsaw, Poland

Email: msyd@ipipan.waw.pl

Abstract—Digital media have enormous impact on the public opinion. In the ideal world the news in public media should be presented in a fair and impartial way. In practice the information presented in digital media is often biased and may distort the opinion on a given entity/event or concept. It is important to work on tools that could support the detection and analysis of the information bias. One of the first steps is to study the methods of automatic detection of the articles reporting on the same topic, event or entity to further use them in comparative analysis or building a test or training set.

In this paper we report on the experimental results concerning the problem of automatic detection of articles reporting on the same events or entities. We also report some experiments on detecting the source of information based on the content.

I. INTRODUCTION

In the ideal world the news in the public media should be presented in a fair and impartial way so that the reader is provided with honest and high-quality unbiased information and can make his own opinion about the political, economical, social or historical events, entities or concepts, etc.

Impartiality of the media that present news to the citizens is a crucial property of democratic system and is what one would expect.

For several reasons the information presented in the media is usually far from being impartial. One of the reasons for this is that various people may see the world events differently what may influence the way they present them. More importantly, in some cases the authors of news articles can intentionally introduce some bias into their publications, e.g. for political reasons, etc.

The problem is even more important in cases when some media (web portals, magazines, etc.) *systematically* introduce consequent intentional bias into the published content in order to intentionally misinform the reader about the state of the world.

This problem is important especially for the digital media, since they have significant influence on public opinion. The way they work changes over the time but they still remain one of the main source of information about daily events. The problem of text bias is common. It happens in newspapers, blogs, social networks etc. Each source of media may represent different point of view. Even such media as news portals, that should present impartial information, can describe events or people framing it differently.

It would be very valuable to work on tools that could support the detection and analysis of such systematic or intentional information bias in digital media in order to contribute to improve the quality and fairness of the information provided to the citizens.

Such tools are very complex and involve interdisciplinary approach including the elements of artificial intelligence, text mining, statistical data analysis, psychology, sociology, etc. One of a basic modules in any bias-analysis tool is a module that makes it possible to automatically or semi-automatically detect *pairs* of news articles (or, more generally: text documents), that report on the same event, topic or entity. Such pairs of articles, where each article comes from a different *source* (e.g. web portal, particular author, etc.) can be further used to make comparison-based analysis towards detecting information bias. The pairs are also necessary to build a training, test or reference set in the case of machine-learning approach to the described research problem.

In this paper we present a method and experimental results of detecting pairs of news articles on the same (similar) topic or reporting the same (similar) event, etc. We focus here on the news articles in news portals.

II. RELATED WORK

There exist multiple approaches to identifying text bias. For most of them the first inevitable phase of bias identification is to find the pairs (or clusters) of similar articles, paragraphs or sentences.

A. News articles similarity

The approach of finding similar texts by using Siamese networks is described in [8]. Siamese networks describe how similar a pair of text documents are. This networks use the same architecture of network and feed two text documents as an input. Then, given such an input pair, an output in the form of the value representing the distance, for example Manhattan distance, between the two text documents from the output is calculated as a measure of (dis)similarity.

In the paper [3] the author describes document text representations and variety of similarity measures for text clustering. They include the measures like: cosine similarity, Euclidean distance, Jaccard coefficient, Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence. Then, based on

the results of a clustering algorithm, there is made a comparison of the results on datasets including variety of news articles, academic papers and web pages.

Authors of the article [12] describe a similarity measure for news recommender systems. In this work there is made a comparison-based analysis of human judgement of similarity and some other measures such as: Lin and WASP measures.

Another work [7] presents research concerning articles on events. In particular it concerns tracking similar articles and clustering them to summarise the events under interest.

B. Media bias

In work [11] the authors identify the news framing which is the way of presentation of news. They compare it to google trends and demonstrate how news framing influences the public attention. They used the concept of mean similarity of a corpus. The mean similarity is calculated on pairs of n articles by average cosine distance of DocVec representation of articles. They discovered that the public opinion change with the mean similarity.

In paper [10] the authors describe some linguistic features that reveal the bias in a text. They refer to the form of verbs, part of speech tags and subjective words. Instead of news articles they used data from Wikipedia, however their results may be used also for other types of texts.

A related work concerning the usability of linguistic features in the task of detecting special form of bias related to the phenomenon of Web Spam is presented in [9].

An interesting approach is presented in the very recent paper [2], where the authors identify three roles of entities framing people in news articles. These roles are hero, villain and victim. The results are presented in a visual form to compare how entities are described in different articles.

Article [4] presents an approach of identifying bias through analysis of mentions and quotations of politicians among different parties and in different periods of time.

Bias and its propagation is also investigated in social networks [6]. This work used twitter data to identify bias in short texts and to analyse its propagation among the users.

III. PROBLEM SPECIFICATION

In this paper we consider two research problems:

- news similarity detection problem
- information source recognition problem

A. News similarity detection

We proposed two approaches to news similarity detection: find all similar articles to the given one and given two articles decide if they are similar or not. The first approach to the problem is as follows. In a given collection of news articles from a given time window (e.g. particular day, etc.) detect the groups of articles that report on the same topic/event, etc. A manually labelled training set that is a collection of manually grouped news articles is prepared. We apply text mining techniques to identify the similar events. The models

are evaluated using the metrics presented in the next section: averaged precision, averaged recall and averaged F-measure.

The second approach is to identify if 2 articles are similar. We apply the machine-learning approach to this problem. For each article there are computed several attributes based on the textual contents, keywords, etc. Then, the set is used to train some ML models. Finally, the models are used to automatically detect similar articles. The models are evaluated using the prepared group labels and some standard measures such as precision, recall or f-measure.

B. Towards news bias detection

The second research problem studied in this paper is the following. Given an article and the set of information sources (e.g. web news portals) is it possible to automatically recognise which source does this article comes from based only on the contents? This kind of experiment can be viewed as a one of simple tests of impartiality of information sources. I.e. if it is possible to correctly predict the source of the news article based on its content then it is more likely that this information source has some information bias.

Of course some other reasons may make it possible to predict the source of the news article including the writing style, etc. However this simple test may serve as one of the multiple tools that could in ensemble help to detect information bias. More advanced bias-detection tools are envisaged in our ongoing research.

IV. EXPERIMENTAL SETUP

A. Data Collection

We collect the data from two Polish news portals: 'dorzeczy.pl' and 'gazeta.pl'. These two are chosen from among the most popular Polish news portals. In addition, they are considered by many readers as examples of media having completely different views on the reality in Poland especially in the domain of social issues or politics and hence making it possible to build an interesting dataset with a potential of containing pairs (or clusters) of articles on the same/similar topic/event/entity but with potentially various forms of information bias. Articles are categorised by a predefined set of topic categories on each of the portals. The decision was made to focus on events connected to the politics in Poland or world. In this case we were looking for articles from category 'Information' ('Wiadomosci') in 'gazeta.pl' and in portal 'dorzeczy.pl' for categories 'Country' and 'World' ('Kraj' and 'Swiat').

In this work we decided to focus only on Polish media but it is possible to extend our research to other languages.

We have collected the articles from 01.01.2018 to 07.04.2018. Table I presents the number of articles.

1) *Database*: We store the data in MongoDB - a document database. We create article collection of news articles and their comments. The comments made by the users to the articles are not used in the experiments reported in this article but are kept for future, extended analyses. Each item in the collection

Table I
DATASET

| news portal | number of articles |
|-------------|--------------------|
| gazeta.pl | 2623 |
| dorzeczy.pl | 4197 |

Table II
DISTRIBUTION OF ARTICLES AMONG GROUPS

| group quantity | number of groups |
|----------------|------------------|
| 1 | 145 |
| 2 | 36 |
| 3 | 17 |
| 4 or more | 16 |

represents an article and contains the following fields: '_id', 'article_id', 'title', 'date', 'lead_text', 'text', 'keywords', 'source', 'url' and 'comments'. Field comments contains 'author', 'date', 'comment_id', 'text'.

B. Data Annotation

For news similarity detection we needed to manually create an annotated data set. The common approach for text similarity recognition is to create a set of article pairs and annotate if they are similar or not. We realised that for news articles this approach may not be the best one. We want to find all articles that are similar and sometimes one news portal describes an event in one article and other news portal writes about this in a series of four articles, for example. Thus we define the task of annotating similar articles as follows.

For a given time window (e.g. a particular date) we collect all articles from the specified web news portals. Each article is assigned to a group with articles about similar event using some particular method. If there is no group with articles describing the event create there is created a new one. The group contains all articles about the same event.

We have annotated 385 articles from 6 randomly chosen days. Articles formed 213 groups. There are groups of consisting of one article or groups containing many articles. The distribution of articles among groups quantity is presented in table II Each record of annotated data contains (among others) the following attributes: 'date', 'article_id', 'group_id'.

C. Data Preprocessing

In the preprocessing phase we apply several operations including: removing stop-words, normalising- convert words to the base form using *Morfeusz* library [1].

D. Evaluation Measures

In order to evaluate experimental results we calculate the average precision, recall and f measure in each experiment.

The average precision is the average of precision of each group. That is given by the following expression:

$$ap = \frac{1}{N} \sum_{n=1}^N p_n = \frac{1}{N} \sum_{n=1}^N \frac{tp_n}{tp_n + fp_n} \quad (1)$$

Where N is the number of evaluated groups. Accordingly average of recall is given as:

Table III
EVALUATION OF ARTICLE'S SIMILARITY DETECTION

| Algorithm | ap | ar | af1 |
|------------------------------|-------------|-------------|-------------|
| Keywords similar. | 0.60 | 0.57 | 0.42 |
| Doc2Vec + cos sim | 0.72 | 0.57 | 0.50 |
| Doc2Vec+bigram+cos similar. | 0.93 | 0.60 | 0.64 |
| Doc2Vec+trigram+cos similar. | 0.92 | 0.63 | 0.66 |
| TF-IDF +cos similar. | 0.50 | 0.69 | 0.53 |

$$ar = \frac{1}{N} \sum_{n=1}^N r_n = \frac{1}{N} \sum_{n=1}^N \frac{tp_n}{tp_n + fn_n} \quad (2)$$

Finally, average F-measure is defined as follows:

$$af1 = \frac{1}{N} \sum_{n=1}^N \frac{2 * p_n * r_n}{p_n + r_n} \quad (3)$$

V. EXPERIMENTAL RESULTS ON ARTICLES SIMILARITY DETECTION

A. Group approach

In a group approach of finding similar articles we experimented with three methods for the news similarity detection problem:

- keyword set similarity - this is our simple baseline solution. We compared the number of similar keywords and find the most similar articles using predefined threshold based on the number of keywords.
- tf-idf with cosine similarity- after preprocessing of a textual data, we calculated tf-idf and cosine similarity between articles from a given data frame. Again we choose the most similar articles based on the predefined threshold.
- doc2vec[5] with cosine similarity in three variants: unigrams, bigram phrases, trigram phrases. For Each of these we choose doc2vec based on bag of words model. Similar preprocessing was done as for tf-idf.

The results of evaluation are presented in a table III. The best averaged results for a given measures were highlighted in bold. The best average precision is observed for two doc2vec models. That means for these models there are the least false positives. However the best f-measure and recall is observed for tf-idf algorithm. This algorithm is better choice if we want to find as many similar articles as possible without caring about dissimilar articles among them.

B. Pair approach

In this task we wished to identify if a pair of articles is similar or not. The data was split into test and train datasets as presented in IV. Similar articles was labelled as '1' and not similar articles as '0'. We have created the following features: cosine similarity on tf-idf vectors, number of similar keywords, normalized number of similar entities that is number of similar entities/sum of entities in both articles. In table V we present an evaluation of proposed algorithms.

Table IV
NUMBER OF ARTICLE PAIRS FOR SIMILARITY DETECTION

| news portal | training set | test set |
|-----------------|--------------|----------|
| similar (1) | 320 | 151 |
| not similar (0) | 7837 | 2624 |

Table V
EVALUATION OF ONE TO ONE ARTICLES PAIRS

| Algorithm | Class | Precision | Recall | F1-score | Support |
|------------------------------|-------------|-----------|--------|----------|---------|
| Siamese LSTM | 0.0 | 0.95 | 0.86 | 0.90 | 2624 |
| | 1.0 | 0.07 | 0.19 | 0.10 | 151 |
| | avg / total | 0.90 | 0.82 | 0.86 | 2775 |
| SVM polynomial kernel | 0.0 | 0.98 | 0.91 | 0.94 | 2624 |
| | 1.0 | 0.29 | 0.63 | 0.40 | 151 |
| | avg / total | 0.94 | 0.90 | 0.91 | 2775 |
| SVM linear kernel | 0.0 | 0.98 | 0.85 | 0.91 | 2624 |
| | 1.0 | 0.20 | 0.68 | 0.31 | 151 |
| | avg / total | 0.94 | 0.84 | 0.88 | 2775 |
| Logistic regression | 0.0 | 0.98 | 0.89 | 0.93 | 2624 |
| | 1.0 | 0.24 | 0.63 | 0.35 | 151 |
| | avg / total | 0.94 | 0.87 | 0.90 | 2775 |
| Gradient boosting classifier | 0.0 | 0.96 | 0.95 | 0.96 | 2624 |
| | 1.0 | 0.27 | 0.28 | 0.27 | 151 |
| | avg / total | 0.92 | 0.92 | 0.92 | 2775 |

Except of Siamese LSTM all algorithms have quite good results. Support vector machines occur to be the best one. Siamese neural networks has high results in total but very low scores for similar pairs where the reason may be that it was not able to detect dependencies in long text.

VI. EXPERIMENTAL RESULTS ON NEWS ARTICLE SOURCE DETECTION

We experimented with three machine learning algorithms in the problem stated as prediction of the news article source based on its content. In all the experiments concerning this problem, the articles' attributes explicitly mentioning the actual source (e.g. the "source" attribute) were ignored in the prediction phase. Dataset for this task is presented in table VI. We have used the following algorithms: naive bayes, logistic regression, support vector machines.

The evaluation of proposed methods is presented in table VII. Support vector machines has the best score slightly outperforming logistic regression. These results show that based on simple approach, analysing the basics of used language we are able recognise the source.

VII. SUMMARY AND FUTURE WORK

Our ongoing research concerns the helper problem of recognizing news articles on (nearly) the same topic/event in order to find media bias. We have proposed 2 approaches and presented their advantages and disadvantages.

Table VI
NUMBER OF ARTICLES FOR MEDIA OUTLET DETECTION

| news portal | training set | test set |
|-------------|--------------|----------|
| gazeta.pl | 2436 | 395 |
| dorzeczy.pl | 3591 | 606 |

Table VII
EVALUATION OF ARTICLE'S MEDIA OUTLETS DETECTION

| algorithm | news portal | precision | recall | f1-score | support |
|---------------------|-------------|-------------|-------------|-------------|---------|
| Naive Bayes | dorzeczy.pl | 0.69 | 0.98 | 0.81 | 606 |
| | gazeta.pl | 0.89 | 0.32 | 0.47 | 395 |
| | avg / total | 0.77 | 0.72 | 0.67 | 1001 |
| Logistic Regression | dorzeczy.pl | 0.90 | 0.83 | 0.86 | 606 |
| | gazeta.pl | 0.76 | 0.86 | 0.81 | 395 |
| | avg / total | 0.85 | 0.84 | 0.84 | 1001 |
| SVM | dorzeczy.pl | 0.88 | 0.86 | 0.87 | 606 |
| | gazeta.pl | 0.79 | 0.83 | 0.81 | 395 |
| | avg / total | 0.85 | 0.85 | 0.85 | 1001 |

We also presented preliminary results on predicting the source of news article based on the contents that seems to illustrate that bias might be present as one of the aspects making such prediction possible, however this needs deeper analysis.

Since some news articles often report multiple events, to improve our results, we plan to increase the granularity of recognition i.e. add recognising fragments of articles instead of the whole documents about similar events. That means that it is planned to detect fragments of text concerning similar events and detect bias in them. Also, we aim to extend research to other languages (e.g. Polish, English).

REFERENCES

- [1] <http://sgjp.pl/morfeusz/morfeusz-siat.html>.
- [2] D. Gomez-Zara, M. Boon, and L. Birnbaum. Who is the hero, the villain, and the victim?: Detection of roles in news articles using natural language techniques. In *23rd International Conference on Intelligent User Interfaces*, pages 311–315. ACM, 2018.
- [3] A. Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, pages 49–56, 2008.
- [4] K. Lazaridou and R. Krestel. Identifying political bias in news articles. *Bulletin of the IEEE TCDL*, 12, 2016.
- [5] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org, 2014.
- [6] H. Lu, J. Caverlee, and W. Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222. ACM, 2015.
- [7] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, pages 280–285. Morgan Kaufmann Publishers Inc., 2002.
- [8] P. Neculoiu, M. Versteegh, and M. Rotaru. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, 2016.
- [9] J. Piskorski, M. Sydow, and D. Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.
- [10] 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)*, volume 1, pages 1650–1659, 2013.
- [11] K. Sheshadri, C.-W. Hang, and M. Singh. The causal link between news framing and legislation. *arXiv preprint arXiv:1802.05768*, 2018.
- [12] N. Tintarev and J. Masthoff. Similarity for news recommender systems. In *Proceedings of the AH&Z06 Workshop on Recommender Systems and Intelligent User Interfaces*. Citeseer, 2006.