Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 178 (2020) 205–212

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Young Scientist Conference on Computational Science (YSC 2020)

Propaganda Identification Using Topic Modelling

Yakunin Kirill[a,c] [0000-0002-7378-9212], Ionescu George Mihail[d] [0000-0002-4531-5298], Murzakhmetov Sanzhar[a,c] [0000-0001-6494-8982], Mussabayev Rustam[c] [0000-0001-7283-5144], Filatova Olga[e] [0000-0001-9568-1002] and Mukhamediev Ravil[a,b,c,*] [0000-0002-3727-043X]

[a]*Satbayev University, 22A Satbayev St., Almaty, Kazakhstan*
[b]*ISMA University, 1 Lomonosov St., Riga, Latvia*
[c]*Institute of Information and Computational Technologies, Pushkin St., Almaty, Kazakhstan*
[d]*University of Bucharest, 90, Panduri Street, Sector 5, Bucharest, ROMANIA*
[e]*St. Petersburg State University, 7 Universitetskaya Emb., St.Petersburg, Russia*

Abstract

This paper presents a method based on topic modelling for identifying texts with propagandistic content. The method is an attempt to incorporate transfer learning idea of obtaining effective vector representation from a large unlabeled or (semi-) automatically labelled dataset, while also attempting to minimize the amount of necessary manual expert labelling by introducing high level labelling (either manual or automatic) on some explicit document property. The proposed method includes four key stages: formation of corpus partitioning, computing a topic model of a united corpus, calculation of corpora imbalance estimates of each topic; extrapolating the results of the imbalance estimation on all documents. The method was cross-validated on a labelled subsample of 1000 news, and achieves high predictive power – ROC AUC 0.73.

Keywords: propaganda; natural language processing; topic modelling; text classification; mass media analysis

1. Introduction

Propaganda has always existed. Its origins can be traced back to the Age of Antiquity. It was effectively used in the Roman Empire, during early Christian era, and in religious conflicts of the Reformation era. It is known that the term “propaganda” itself began to be used in 1622 (when the pope founded the “The Congregation for the Propagation of the Faith”), it gained popularity during the First and Second World Wars, but the greatest popularity was achieved during the Cold War [1, 2].

Corresponding author e-*mail address:* ravil.muhamedyev@gmail.com

10.1016/j.procs.2020.11.022

However, research on propaganda is becoming especially relevant right now, when propaganda is becoming more "digital". Moreover, unlike the First and Second World Wars, now we are dealing with local conflicts, and, unlike the Cold War period, people from both sides have full access to the media of the other side [2].

By political propaganda, we mean the coordinated, systematic informational influence of the propaganda subject on target audiences in order to achieve political goals and promote political ideas.

The increasing use of fake news, misinformation, and propaganda is concerning international media freedom observers. So, in 2015, the OSCE Special Representative on Freedom of the Media prepared a report entitled "Propaganda and freedom of the media" [3], which indicates that it is necessary to distinguish between two types of propaganda in the media. The first is propaganda for war and inciting national, racial or religious hatred, which constitutes an incitement to discrimination or violence, as defined in international and national laws. This is an illegal form of propaganda. The second type includes all other types of propaganda. Moreover, the Representative believes that propaganda spread through state-owned media, controlled by the state or by its representatives is especially dangerous.

In modern conditions of increasing information confrontation between countries, the propaganda and counter-propaganda come to the forefront, as states need to protect their citizens from various information threats and ensure their security, which is a prerequisite for the further development of the state. In order to achieve this studies that test methods for identifying propaganda among a large volume of documents are necessary.

The main goal of this article is to demonstrate the possibility of using an approach based on topic modelling to identify propaganda in the media.

In Related works, we briefly review the available results of propaganda analysis and automatic text processing.

Data and method part describes the applied methods of topic modelling and the used data.

In Results and discussion part the results are described and discussed.

In conclusion, we summarize the discussion and formulate the directions for further studies.

In our opinion, the described approach for identifying such a semantically fuzzy phenomenon as propaganda without conducting a large-scale labelling of documents or parts of them is proposed for the first time.

2. Related works

Natural language processing (natural language processing – NLP) is an area of research concerned with automatic analysis of big volumes of textual data. These technologies allow users to obtain information from large amounts of textual data [4], provide content analysis [5], personalized access to news [6] and even help their production and dissemination [7]. According to [8], the key aspects that allowed us to obtain impressive results in the field of automatic natural language processing are the novelties in the development of machine learning methods, orders of magnitude increases in computing power, a large amount of linguistic data and the development of an understanding of the natural language structure applied to the social context.

The problem of classifying text's propagandistic content was addressed by several researches [9, 10]. However, the amount of research is orders of magnitude smaller, than in the area of sentiment analysis. For example, [11] presents a bag-of-words based approach, working on the level of sentences. It was compared to BERT as a baseline. The achieved result is classification with F1 Score ~0.6.

At the same time, one of the methods productively applied in the field of NLP is topic modelling. Topic modelling is a method based on the statistical characteristics of document collections, which is used in the tasks of automatic abstracting, information extraction and classification [12]. The essence of this approach is to find groups/clusters/topics of documents in a corpus in which the frequency of words or word combinations cooccurrence is high. Probabilistic topic models describe documents by a discrete distribution on a variety of topics, and topics by a discrete distribution on a variety of terms [13]. In other words, the topic model determines which topics each document relates to and which words form each topic. Clusters of words and phrases formed in the process of topic modelling in particular, allows solving problems of terms' synonymy and polysemy [14].

In this paper, we propose an approach to identify texts containing propaganda using a topic model of the text corpus. The proposed approach differs from the above presented research in that it attempts to perform analysis on a much higher level of abstraction (topics and relation of texts to topics, rather than separate word in a sentence). However, proposed approach requires presence of some explicit property for partitioning (for example news source) and big volume of corpus (at least hundreds of thousands of documents), which also differentiates it from other research in the field.

Another area of research worthy of mention in the context of identifying propaganda in mass media is identifying bot activity in mass media and social networks. Notably, there are a number of tools for analyzing and identifying bot activity, such as Hamilton 2.0 Dashboard by Alliance for Securing Democracy [15], Botometer by the Network Science Institute and the Center for Complex Networks and Systems Research [16]; and NATO Strategic Communications Centre of Excellence publishes regular reports on social media bots activity titled Robotrolling [17].

3. Data and method

The rationale behind the proposed model is that classical approach to text classification requires substantial amount of manual text labelling. While for more researched areas such as sentiment analysis massive corpora of labelled texts exist, for less researched problems, such as propaganda, social importance, resonance/popularity classification and others, this is not the case. Since in social sciences, politics research and other humanitarian areas there is a demand for various text classification problems, not limited by sentiment analysis, we propose a model which attempts to build classifiers based on minimal manual expert labelling (in this case – labelling of sources) or without any labelling at all. The latter is the case if there is some explicit property, which allows partitioning the corpora such that it tightly corresponds with the target implicit property. For example, the problem of news popularity/resonance – it is tightly related to number of views, which is an objective parameter, so one can propose a model to partition a corpus based on the property automatically.

The proposed model can also be viewed as an alternative approach to transfer learning, in that it attempts to obtain effective vector representation (topic embedding) from a big volume of unlabeled data. Hence, it should be noted, that even those documents, which cannot be assigned to any particular corpus/property, can still be used during the topic modelling phase, in order to obtain more effective vector representations.

The proposed method includes four stages:

- Formation of corpus and partition into separate corpora based on some explicit (objective) property (in our case – news source)
- Computing a topic model of a united corpus
- Estimation of corpora imbalance of each topic of the obtained topic model
- Extrapolating the results of the imbalance estimation on all documents/news, even those which were not present in the original corpora (for example if the degree of bias/propaganda of a certain source is unknown or controversial).

3.1. Stage 1. Formation of the corpuses

For the problem of identification of propagandistic news articles, a corpus based on free open Russian news websites was formed. The proposed method implies that the corpus should be divided into two or more separate corpuses based on some explicit property (in our case – news source), in order to attempt to reveal some features in accordance with some implicit property (in our case – degree to which each news publication can be considered to be propagandistic). Hence, based on consideration, discussed below, the corpus of 428180 news articles published in 2018-2020 was divided into two corpora based on their source:

- Propaganda corpus (346440 documents):
    - Russia Today
    - Current Time
    - Radio Liberty
    - Deutsche Welle
    - Sputnik
- Rhetorically impartial/Non-biased corpus (81740 documents):
    - Vedomosti
    - Interfax

- Lenta.ru
- Business FM
- RBC

This distinction was based on a number of considerations. The Russian language mass-media landscape in Russian Federation consists of state-owned governmental holdings that are controlling the activity of the media, private or independent mass-media (so called opposition or liberal media in Russian), and foreign public diplomacy mass-media, which are broadcasting in Russian language, are owned by the governments of United States, Great Britain, Germany and France, and are registered in Russia as a foreign agent according to the Federal Law no. 426-FZ from 02.12.2019. There are also two governmental media, Sputnik and RT, that are involved in Russia's public diplomacy, but due to internal regulations they do not broadcast to internal Russian public, but to Russian speaking CIS public. For a more concluding study we decided to analyze propagandistic media that have a clearly subjective rhetoric compared to non-propagandistic media, which rhetoric is as objective as possible.

The international broadcasters are part of the state public diplomacy machine and they are clearly propagandistic. Their objective is to promote the state image and political interests to foreign publics. Thus, we analyzed such media as RT, Sputnik, Radio Free Europe/Radio Liberty, Current Time, and DW

On the other hand, not all private Russian mass media are propaganda free. If the state media adopted a positive rhetoric regarding the governing political class, the so-called opposition media are directly criticizing the government. Therefore, we excluded from our analysis such private Russian mass media as TV Rain, Meduza or Novaya Gazeta due to anti governmental propaganda. Instead we decided to take into account such Russian media that are less involved directly in politics and more involved in business or economics as such media have a more objective rhetoric. Hence, we chose to analyze two quite neutral news agencies (Interfax), three business-oriented media (RBC, Vedomosti, and Business FM), and one generalist online newspaper (Lenta.ru).

3.2. Stage 2. Topic modelling

In order to build a topic model of the corpus of documents, there are used: Probabilistic latent-semantic analysis (PLSA), ARTM (Additive regularization of topic models) [18] and the very popular latent Dirichlet allocation (LDA) [19, 20]. LDA can be expressed by the following equality:

$$p(w,m) = \sum_{t \in T} p(w \mid t,m)\, p(t \mid m) = \sum_{t \in T} p(w \mid t)\, p\,(t \mid m) = \sum_{t \in T} \varphi_{wt}\theta_{tm} \tag{1}$$

representing the sum of mixed conditional distributions over all topics of the set T, where p (w | t) is the conditional distribution of words in topics, p(t | m) is the conditional distribution of topics on the news. The transition from the conditional distribution p(w | t, m) to p(w | t) is carried out due to the hypothesis of conditional independence, according to which the appearance of words in the news m on the topic t depends on the topic, but does not depend on the news m, and it is a common for all the news. This ratio is true, based on the assumptions that there is no need to preserve the order of documents (news) in the corpus and the order of words in the news. In addition, the LDA method assumes that the components φ_{wt} and θ_{tm} are generated by a continuous multidimensional probability Dirichlet distribution.

The purpose of the algorithm is to search for the parameters φ_{wt} and θ_{tm}, by maximizing the likelihood function with the corresponding regularization

$$\sum_{m \in M} \sum_{w \in m} n_{mw} ln \sum_{t \in T} \varphi_{wt}\theta_{tm} + R(\varphi,\theta) \rightarrow max \tag{2}$$

where $\boldsymbol{n_{mw}}$ is the number of occurrences of the word w, in the news m, R(φ,θ) is the logarithmic regularizer. To determine the optimal number of thematic clusters T, the method of maximizing the coherence value calculated using the UMass metric is often used [21].

The results described in the paper were obtained by applying BigARTM topic model with smooth "sparse theta regularizer" (tau=0.15), "decorrelator phi regularizer" (tau=0.5) and "improve coherence phi regularizer" (tau=0.2), the number of topics is 200. The parameters were chosen in the course of experiments.

3.3. Stage 3. Estimation of corpora imbalance

Next step is to estimate the imbalance in the distribution of news of different corpora within each topic. This measure of imbalance will be an estimation of influence of belonging of the given topic to the target implicit property (in this case – propaganda), since the partition of the original corpus was performed based on explicit property (source) while taking into account consideration discussed above, which allow us to argue that there is an inherent imbalance of propaganda/bias in this two corpora. Formula for imbalance estimation:

$$D_{t_i c_j} = \frac{\sum_k w_{d_k t_i c_j}}{\sum_k \sum_l w_{d_k t_l c_j}} \Big/ \sum_m \sum_k \sum_l w_{d_k t_l c_m} \tag{3}$$

where $D_{t_i c_j}$ is a measure to imbalance the presence of documents from the corpus $c_j$in topic t_i, a $w_{d_k t_l c_m}$- document weight $d_k$from corpus $c_m$to topic t_l.

3.4. Stage 4. Extrapolating the results

The last step in the proposed method is to apply the obtained topic model (theta matrix representing the relation of each document to each topic) and imbalance estimations in order to classify each individual document (news article). The reasons for that are:

- Even though the most distinctive propagandistic and unbiased sources were selected, the distribution of propaganda among individual news is certainly uneven, in that some proportion of propaganda can be published in unbiased sources and vice versa.
- As was discussed above, by far not all sources can be assigned to one of the corpora, since the degree of propaganda in some news sources may be uncertain and debatable.

In order to aggregate the imbalance estimation with topic relation vector of each document, several approaches can be applied:

- Simple weighted average – the average of multiplications of document's topics' weights by corresponding imbalance estimates.
- Bayesian approach – this approach considers subjective probability of a document to correspond to the given criterion. Its advantages were described in [22, 23].
- Semi-supervised approach. It is possible to pre-train supervised model on results obtained by approaches a, b or some other unsupervised approach, and then to use document labelling in order to fine-tune the model, thus improving its performance.

The simple weighted average was used to obtain the results described in the next section. It was selected due to two reasons: it is easier to implement and other approaches are considered to be directions of future research, and also it illustrates that the method work adequately even in the case of very simple aggregation method without any trainable parameters.

4. Results and discussion

In order to validate the proposed model, a random representative (by date, sources and topics distributions) sub-sample of 1000 news articles was obtained from the original corpora, which was excluded from the imbalance estimation (training/fitting of the model), and manually labelled by an expert in the field according to Likert scale, were +2 corresponds to strong opinion, that the article contains certain propaganda, -2 corresponds to strong opinion, that the article is unbiased/objective. +1/-1 is, respectively, less strong opinion on propaganda/objectivity of an article, and 0 refers to debatable, uncertain articles.

Then, the imbalance estimates were applied to this subsample, in order to calculate classification quality metrics. Likert scale was linearly normalized to range [0, 1], and extrapolated news estimates were also scaled to [0, 1]. Based on this scaling Pearson correlation and mean absolute error (MAE) were calculated. Then classes were assigned as follows – values more than or equal to 0.5 corresponds to propaganda (positive class), values less 0.5 correspond to unbiased/objective news (negative class). This data was used to obtain F1-Score and ROC AUC. The following results were obtained:

Table 1. Results of model verification.

Classification threshold	Pearson correlation	MAE	F1-Score (micro)	ROC AUC
Without threshold	0.47	0.27	0.72	0.73
<0.4 | >0.6	0.61	0.26	0.81	0.8
<0.3 | >0.7	0.81	0.18	0.94	0.95

In table 1 classification thresholds corresponds to model output values, which were considered confident. For example, <0.4 | >0.6 means that all values in the interval [0.4, 0.6] are considered to be "unknown class" or not confident enough.

Even without such threshold, the model demonstrates considerable predictive power (ROC AUC 0.73), while at <0.3 | >0.7 threshold the predictive power is very high (ROC AUC 0.95), which means that on extreme values the model is very accurate, while it also identifies those news, for which classification of propaganda is not certain or debatable.

5. Conclusion

The problem with studying propaganda, misinformation, and fake news today is that they take many forms and spread through many channels. Communications are developing rapidly and, although propaganda has a long history, its reach, scale and effectiveness, today it has risen to unprecedented levels thanks to the viral speed of social networks and the capabilities of artificial intelligence.

We dare to assert that two epochs can be distinguished in the study of propaganda. The first is conventionally called the era of classical propaganda, and the second - the era of "propaganda 2.0", or "digital propaganda". And if classical propaganda is well studied, its types, goals, channels, audiences, mechanisms and principles of influence are identified, models for its study are built, the "propaganda 2.0" requires the use of modern technologies for automatic text processing.

The proposed model attempts to provide high predictive power, while maintaining necessary volume of manual expert labelling as low as possible. It proposes high-level labelling of corpus according to some explicit property, which must have some correlation with the target implicit property. In this case partitioning by source was used, but other partitioning, including automatic can also be performed depending on target.

The model was cross-validated on a labelled subsample of 1000 news, and shows high predictive power – ROC AUC 0.73 without confidence threshold, and ROC AUC 0.95 with [0.3; 0.7] classification threshold.

The conducted research showed that assessing the level of propaganda is quite difficult - both for experts and for automatic models. Even if there are striking examples of propagandistic content, the deliberately "propagandistic" media often write not only propaganda articles and, on the other hand, not all unbiased/neutral media are always objective.

The method used by us is limited by the applied corpus of documents, the quality of the topic model and expert labelling. At the same time, it can significantly reduce experts' efforts in identifying propagandistic content.

Further research includes applying this model to other text classification problems, and applying other aggregations methods, discussed in the work, such as Bayesian aggregation and semi-supervised approach to aggregation.

Acknowledgements

The work was funded by grant No. BR05236839 of the Ministry of Education and Science of the Republic of Kazakhstan.

References

[1] Garth J, O'Donnell V. (1999). Propaganda and Persuasion. SAGE Publications, Inc.

[2] Filatova O. (2018). Propaganda in the epoch of bots, trolls and fake news: theoretical approaches and applied research. // Strategic communications in business and politics. S.-Petersburg State University. 4: 86-94.

[3] Propaganda and freedom of the media. Memorandum of the Office of the OSCE Representative on Freedom of the Media. Vienna, 2015: 34.

[4] Korenčić, Damir & Ristov, Strahil & Šnajder, Jan. (2018). Document-based Topic Coherence Measures for News Media Text. Expert Systems with Applications. 114. 10.1016/j.eswa.2018.07.063. Vol. 114: 357-373.

[5] Neuendorf, Kimberly. (2016). The Content Analysis Guidebook.// Flaounas, Ilias & Lansdall-Welfare, Thomas & Bie, Tijl & Mosdell, Nick & Lewis, Justin & Cristianini, Nello. (2013). Research methods in the age of digital journalism. Digital Journalism. 1. 102-116. 10.1080/21670811.2012.714928.

[6] Steinberger, Josef & Ebrahim, Mohamed & Ehrmann, Maud & Hürriyetoglu, Ali & Kabadjov, Mijail & Lenkova, Polina & Steinberger, Ralf & Tanev, Hristo & Vázquez, Silvia & Zavarella, Vanni. (2012). Creating sentiment dictionaries via triangulation. Decision Support Systems. 53. 689–694. 10.1016/j.dss.2012.05.029. // Vossen P., Rigau G., Serafini L., Stouten P., Irving F., Van Hage W. R. NewsReader: recording history from daily news streams. In LREC. (2014):2000–2007.// Li, Lei & Zheng, Li & Yang, Fan & Li, Tao. (2014). Modelling and broadening temporal user interest in personalized news recommendation. Expert Systems with Applications. 41. 3168–3177. 10.1016/j.eswa.2013.11.020.

[7] Clerwall, Christer. (2014). Enter the Robot Journalist: Users' Perceptions of Automated Content. Journalism Practice. 8. 519-531. // Popescu O., Strapparava C. *Natural Language Processing meets Journalism*. Proceedings of the 2017 EMNLP Workshop. Copenhagen, Denmark: Association for Computational Linguistics (2017).

[8] Hirschberg, Julia & Manning, Christopher. (2015). Advances in natural language processing. Science (New York, N.Y.). 349. 261-266. 10.1126/science.aaa8685.

[9] Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019, November). Fine-grained analysis of propaganda in news article. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): 5640-5650.

[10] Barrón-Cedeno, A., Jaradat, I., Da San Martino, G.,& Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5): 1849-1864.

[11] Martino, Giovanni & Yu, Seunghak & Barrón-Cedeño, Alberto & Petrov, Rostislav & Nakov, Preslav. (2019). Fine-Grained Analysis of Propaganda in News Articles: 5640-5650)].

[12] Mashechkin I.V., Petrovskiy M.I., Tsarev D.V. Methods of text fragment relevance estimation based on the topic model analysis in the text summarization problem. Computing methods and programming. (2013) №. 1: 91-102.

[13] Vorontsov K.V., Potapenko A.A. Regularization, robustness and sparsity of probabilistic topic models. Computer Research and Modelling, 2012, vol. 4, no. 4: 693-706.

[14] Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A. A survey and an experimental comparison of methods for text clustering: application to scientific articles. Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS). 2017;29(2):161-200. (In Russ.) https://doi.org/10.15514/ISPRAS-2017-29(2)-6

[15] Hamilton 2.0 Dashboard https://securingdemocracy.gmfus.org/hamilton-dashboard/.

[16] Botometer https://botometer.iuni.iu.edu.

[17] Robotrolling (2020/2) https://www.stratcomcoe.org/robotrolling-20202

[18] Vorontsov, Konstantin & Frei, Oleksandr & Apishev, Murat & Romov, Peter & Dudarenko, Marina. (2015). BigARTM: Open Source Library for Regularized Multimodal Topic Modelling of Large Collections. 370-381. 10.1007/978-3-319-26123-2_36.

[19] Blei, David & Ng, Andrew & Jordan, Michael. (2013). Latent Dirichlet Allocation. Journal of Machine Learning Research. 3: 993-1022.

[20] Jelodar, Hamed & Wang, Yongli & Yuan, Chi & Feng, Xia & Jiang, Xiahui & Li, Yanchao & Zhao, Liang. (2018). Latent Dirichlet allocation (LDA) and topic modelling: models, applications, a survey. Multimedia Tools and Applications. 78. 10.1007/s11042-018-6894-4: 1-43.

[21] Mimno, David & Wallach, Hanna & Talley, Edmund & Leenders, Miriam & Mccallum, Andrew. (2011). Optimizing Semantic Coherence in Topic Models. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 262-272.

[22] Barakhnin, V & Muhamedyev, Ravil & Mussabayev, Rustam & Kozhemyakina, Yu & Issayeva, A & Kuchin, Yan & Murzakhmetov, S & Yakunin, Kirill. (2019). Methods to identify the destructive information Methods to identify the destructive information. Journal of Physics Conference Series. 1405. 12004. 10.1088/1742-6596/1405/1/012004.

[23] Mukhamediev, R. I., Mustakayev, R., Yakunin, K., Kiseleva, S., Gopejenko, V.(2019). Multi-Criteria Spatial Decision Making Support system for Renewable Energy Development in Kazakhstan.