Classification of negative publication in mass media using topic modeling

Kirill Yakunin\textsuperscript{1,3}, Ravil Mukhamediev\textsuperscript{1,2}, Yan Kuchin\textsuperscript{1,3}, Rustam Musabayev\textsuperscript{3}, Timur Buldybayev\textsuperscript{4} and Sanzhar Murzakhmetov\textsuperscript{3}

\textsuperscript{1} Satbayev University (KazNRTU), Kazakhstan, Almaty, Satpayev str., 22A
\textsuperscript{2} ISMA university, Latvia, Riga, Lomonosov str., 1
\textsuperscript{3} Institute of Information and Computational Technologies. Kazakhstan, Almaty, Pushkin str., 125
\textsuperscript{4} Information-Analytical Center. Kazakhstan, Nur-Sultan, Dostyk str., 18

e-mail: yakunin.k@mail.ru

Abstract. The paper proposes a method for evaluating text documents by arbitrary criteria, combining the topic modeling on the text corpora and multiple-criteria decision making. The evaluation is based on an analysis of the corpora as follows: the conditional probability distribution of media by topics, properties and classes is calculated after the formation of the topic model of the corpora. Weights assigned by experts to each topic along with topic model can be applied to evaluate each document in the corpora according to each of the considered criteria and classes. The proposed method was applied to a corpus of 804,829 news publications from 40 Kazakhstani sources published from 01.01.2018 to 31.12.2019, in order to classify negative information on socially significant topics. A BigARTM model was calculated (200 topics) and the proposed model was applied. Experiments confirm the general possibility of evaluating the sentiment of publications using the topic model of the text corpora, since ROC AUC score of 0.93 was achieved on the classification task.

1. Introduction

It is important to understand how the media use their influence in order to mitigate the negative impacts of the media and encourage positive effects [1].

Researchers focus their efforts on media content evaluation due to its practical relevance for news agencies, advertising companies, and the public sector. Media content analysis allows us to predict probable popularity of news articles [2], provision of targeted and high-quality articles for users, PR strategies planning to promote goods or services [3, 4]. The public sector receives a tool for promotion and outreach on innovations, PR planning, and identifying negative content prohibited by law. Individual users can quickly and efficiently filter large amounts of information.

The method we proposed is based on topic modeling and Bayesian rules, which reduces the cost of labeling up the corpus of documents. At the same time, the method shows results comparable to more labor-intensive approaches based on supervised learning.

Besides the introduction, the paper consists of the following sections:

The second section describes current problems, mass media monitoring services, their features and limitations.
The third section describes the proposed model for texts evaluation
The fourth section describes the corpora that was processed, properties and classes, and results of
method verification on the task of identification of negative information on significant topics
In conclusion, we discuss the results obtained, the advantages and disadvantages of the proposed
algorithm, and outline the directions for future research.

2. Related works, tools and methods

2.1. Media monitoring tasks
We generalize the definition of media monitoring as process of content analysis of mass media. Media
monitoring can be used for comparison between big corpora of texts (task 1). For example, in [5] a
task of comparison between Turkish and English corpora of news related to science is considered.
Moreover, media monitoring includes such tasks as social behavior, public opinion identification [6, 7],
online sales trend analysis [8], etc. (task 2) comparison of preferences and characteristics of
population segments (task 3). For example [9] analyses gender inequality. Media monitoring tools are
an important part of reputation management (task 4). The most frequently implemented functions:
brand mentions (with or without direct tagging), relevant hash tags (branded and unbranded), mentions
of competitors, general trends applicable to some industry. The main question that can be answered
with the help of such tools: “Do you know what they say about your company or brand in the media?”

In popular systems, mostly reputation management-related tasks are implemented, along with
manual methods of analysis, such as keywords-based queries with application of TF-IDF indicators [7, 10].

Although application of keyword queries provides a certain level of interpretability, it still imposes
some limitations on this online systems and services:
1. As a rule, they are limited in query semantics, query results need further manual selection [5].
2. Results of queries execution depend on search algorithm and current system/database state. Such
services generally do not have features for saving queries results [10].
3. There tools are limited in evaluation criteria that are assessed, usually only sentiment analysis
and some form of media coverage index are assessed.

For more research oriented tasks (comparison of corpora related to different countries/cultural
spaces, population segments preferences and public opinion analysis, topical trends and seasonality,
etc.) automatic clusterization of texts is often performed with application of Latent Dirichlet
distribution (LDA) [11,12], and classification of documents is performed using machine learning
models [12,13]. It should be noted that supervised learning approach is possible if a labelled dataset of
significant volume is available.

In order to solve these tasks in a situation when promptly obtaining big volumes of labelled texts is
not possible, proposed simple evaluation approach can be applied in combination with topic modelling
of news corpora to obtain effective topic-based vector representations.

2.2. Topic modeling
Topic analysis or topic modeling is one of the actively applied NLP methods. Topic modeling is a
method based on the statistical characteristics of collections of documents, which is used for text
summarization, information extraction, information retrieval, and classification [14]. The essence of
this approach lies in the intuitive understanding that the documents in the collection form groups
where the frequency of words or word combinations differs.

Topic modelling is used as an attempt to overcome the limitations of BERT [15] and other deep
learning NLP models such as necessity for big volumes of manual labelling and significant
computational complexity, while also preserving the idea of transfer learning by transferring
knowledge from big unlabeled dataset in order to obtain knowledge on high-level hidden latent
structures of the corpora.
Topic modeling is well developed in terms of algorithms and methods based on a statistical language model, and the use of document clusters related to a set of topics allows solving problems of synonymy and polysemy of terms [16]. Probabilistic topic models describe documents (M) by discrete distribution on the set of topics (T), and topics by discrete distribution on the set of terms [17]. In other words, the topic model determines which topic each document belongs to and which words form each topic. Probabilistic latent semantic analysis (PLSA), a quite popular latent Dirichlet distribution (LDA) methods [18] and its extension BigARTM [19] are used to build a topic model of document corpora.

In this paper, BigARTM topic model is used to calculate conditional probability distributions of documents by topics, properties and one or more classes. The aggregation of topics effect on the final assessment is carried out using the Bayesian method, which ultimately allows calculating the expected hypotheses on certain media publications characteristics.

3. Method
The proposed model considers assessments of each publications in three modalities:
- Topics, obtained through topic modelling. For example, sports, education, economics, accidents, etc.
- Properties – an arbitrary list of properties can be used. It is necessary to have a process to obtain assessments of influence of each topic on each of the chosen properties. Examples of such processes is described in section 4. Examples of properties are sentiment, social significance, objectivity, manipulativeness, politization, etc.
- Classes. Identification of articles related to each class is in general the final goal of the proposed method. The chosen properties should have some correlation with or influence on the final classes. In the case of the article the final class is negative information on socially significant topics.

It should be noted that in this paper only one property, negative sentiment, is considered. Other properties and composite class will be considered in further research.

The main idea of multi-criteria assessment and aggregation of subjective and objective properties was initially described in [20].

Application of topic model generated through cluster analysis (unsupervised learning) in conjunction with classes and attributes defined by experts is noteworthy about the proposed approach. This way, the semantics of the distribution is determined by the user (expert), and the initial topic modeling depends on the corpora of documents.

Application of the Bayesian approach allows assessing the probability of a hypothesis based on incomplete information with a part of the text corpora. In other words, it allows obtaining evaluation estimates for a media for the mentioned modalities by processing only a part of the text corpora, albeit with reduced accuracy.

The purpose of the algorithm to aggregates the weights of the correspondence of articles to topics, and then the correspondence of topics to parameters/properties and classes. This allows obtaining correspondence estimates of each document in three modalities: topics, properties, and classes. We named the algorithm as multi modal mass media assessment (M4A).

Firstly, we generate the topic model and then calculate the values of conditional probabilities $p_1, ..., p_6$ (see further).

Let us describe the process of how the M4A model is calculated. There are several computational details worth mentioning first in the context of Bayesian aggregation.

If subjective probability $p(e|h)$ is equal to 0.5, it will not have an impact on the target hypothesis $p(h|e)$ and can be interpreted as inconclusive or irrelevant. Hence, if $p(e|h)$ is higher than 0.5, it’s effect on target hypothesis is positive, and if it’s lower than 0.5 – negative.

This peculiarity implies that the normalization at each step of the calculation process should be adjusted correspondingly. For example, if a feature/event has a positive impact 0.8, but it is only applied to a certain object with weight 0.5, it is not correct to just multiply this two numbers and then
renormalize it, since the result of multiplication would be 0.4, which is a small negative impact, which does not conform with actual feature/event impact of 0.8 (positive). And event if one would attempt to later renormalize the results, there is not guarantee, that the middle (0.5) would be kept intact.

It should also be clarified, that the weight (0.5) in the previous example usually corresponds to the degree to which document corresponds to a topic, which was assigned a certain impact value (0.8). However, that may change from one step of the model to another as will be noted below.

In order to resolve this issue, a custom process of normalizing the weighted impact was introduced, which is performed according to the following formula:

\[(p - 0.5) * w + 0.5\]  (1)

where \(p\) is the impact of a feature/event/topic and \(w\) is the weight (or the degree to which the object relates to the given feature/event/topic). After that, the Bayesian aggregation process described above is applied. The last step is the normalization of each row of the resulting matrix, which is also customized – all values below 0.5 are normalized separately to the range [0; 0.5] and all values above 0.5 are normalized to [0.5; 1]. This normalization allows preventing values damping in order to keep values at adequate level of saturation along the multiple matrix transformations (p1->p6).

Now let us described the process of calculation of P matrices:

- P1 matrix describes the relation between topics and evaluation criteria. It can be obtained in different ways, including dictionary generation, manual labeling and (semi-) automatic labeling using multi-corpus approach [21].
- P2 matrix describes the relation between the documents and the topics and is obtained through topic modelling (in LDA-based topic models it corresponds to theta matrix).
- P3 matrix is obtained from performing first stage of AHP – reducing pairwise comparison matrix of evaluations criteria to a column vector of each criterion importance.
- P4 describes the relation between topics and target classes. P4 matrix is calculated as a custom matrix multiplication of p1 by p3. By custom matrix multiplication we refer to the custom weighted impact with customized normalization, as described above.
- P5 describes evaluations of relations of each document to each target class and is calculated using the custom matrix multiplication.
- P6 describes evaluations of relations of each document to each evaluation criterion, and is also calculated using the custom matrix multiplication.

The main result of the whole process is P5 matrix, which describes the relation of each document to target classes, which are composed of evaluation criteria.

P6 matrix can also have some utility based on how independently useful the evaluation criteria are.

4. Data and verification

P2 matrix was obtained through BigARTM topic modeling on a corpus of 804829 news publications from 40 Kazakhstani sources published from 01.01.2018 to 31.12.2019. Corpus and data used for verification is available at [https://github.com/KindYAK/M4A-Data].

BigARTM model was calculated 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 grid search optimization.

Negative sentiment values for each topic were obtained by manual labelling of two experts, which were cross-checked for inconsistencies, corrected and averaged. Each topic was represented by 25 top word/phrases and each expert was asked to give an estimate of this topic’s sentiment on a scale from 0 to 10, were 5 is neutral, 0 is very positive and 10 is very negative. It should be noted, that the sentiment labelling may not reflect author’s opinion in each text, by rather the negativity or positivity of an event described in a given news publication.

Let us consider the top news by negative sentiment:
Table 1 shows top 5 news by negative sentiment. Top news by negative sentiment consist mostly of homicide, planned terrorist activities investigation, catastrophes (such as Arys explosion in 2019) and human rights related issues (rallies, freedom of assembly, police brutality, etc.).

The table illustrates that top news by estimated values indeed correspond to the respective criteria. However, such manual verification naturally cannot be used to verify model’s results. Hence, a methodology for cross-validation of news was proposed:

1) A subset of 10% of news publications were randomly labelled as test set
2) The described model calculation was performed, starting from topic modelling stage, without taking the news in test set into account
3) Then when all necessary weights were obtained (matrices $P_1$-$P_4$), matrices $P_5$ and $P_6$ were calculated for the whole corpus, both test set and training set.
4) A random subsample of 1000 news labelled as test set were randomly selected from top 10 percentile and bottom 10 percentile according to calculated estimates. The reason behind considering only top and bottom news is that experiments have shown that a significant proportion of news cannot be definitely labelled by neither expert, nor the system, since not all news for example are definitely positive or negative, especially since we are dealing with professional journalism which is supposed to be neutral and unbiased. It should be noted, that 20 percent of the corpora is still a considerable volume of news – around 160000. Hence, the goal of the verification is to verify that precision of the model is high, while recall is not considered to be critical according to the proposed methodology.
5) These subsamples were manually labelled by experts. Experts were provided with title, date, url and media source.
6) Quality metrics were calculated

Table 2. Model verification results

| Criterion       | MAE | F1-Score | ROC AUC |
|-----------------|-----|----------|---------|
| Negative Sentiment | 0.14 | 0.93     | 0.93    |

Table 2 shows the results of model verification. Negative sentiment was predicted with ROC AUC of 0.93, with precision of 0.94 and recall of 0.94. It should be noted, that such results were obtained with relatively simple model (without any elements of RNN or Deep Learning) and all necessary weights/parameters calculation required minimal manual labelling.
5. Conclusion
The paper proposes a method allowing classifying publication by any given criterion.

The proposed model was validated both based on lists of top news publications by negative sentiment, which looks coherent and logically consistent, and based on standard ML metrics such as F1 Score and ROC AUC calculated on subsample of 1000 news manually labeled by experts.

However, the described approach has a number of limitations:

- The topic modeling applied to obtain vector representations of texts performs analysis on so-called bag of words level and does not take into account more subtle semantics of the texts, order of words an n-grams, etc., which considerably limits potential quality of the algorithm.

- Expert assessments of the significance of attributes inherit a large share of subjectivity.

- The described approach performs best on big corpora (at least 50-100 thousand texts) of long texts (at least 100-150 words, preferable longer). Preliminary experiments have shown, that applying the proposed model to smaller corpora, especially if it’s not representative of the general population, the model shows much lower predictive power.

To overcome these shortcomings, further studies suggest the application of models using distributive semantics methods to form a topic model, or to move the problem of high-level labelling to text embedding space. Such an approach would presumably take into account the semantic content of the text, in contrast to counting the number of words.

The proposed model was implemented as a scheduled worker in informational system for mass media monitoring and evaluation [22]. The source code of the system is available at [https://github.com/KindYAK/NLPMonitor], source code for workers (preprocessors, models, parsers) is available in a separate repository [https://github.com/KindYAK/NLPMonitor-DAGs]. Other similar systems mentioned in the paper are mostly aimed to monitor certain brands or products in mass media, and perform mostly statistical analysis by keyword searches and forming statistical reports on word, phrases and tags frequency, list of publications by date/popularity/source, etc. The developed system with the proposed model integrated allows solving both classical problems such as simple reports or sentiment analysis, but also has a number of unique use-cases, which distinguish it from existing solutions, such as topic-level analysis and ability to perform analysis of entities without keyword-based search.

Funding Statement
The work was funded by a grant BR05236839 of the Ministry of Education and Science of the Republic of Kazakhstan.

References
[1] Bushman B and Whitaker J 2017 Media Influence on Behavior Reference Module in Neuroscience and Biobehavioral Psychology
[2] Mishra S, Rizoiu M A and Xie L 2016 Feature driven and point process approaches for popularity prediction Proceedings of the 25th ACM international on conference on information and knowledge management pp 1069–78
[3] Tatar A, Antoniadis P, Amorim M D D and Fdida S 2012 Ranking News Articles Based on Popularity Prediction 2012 IEEE/ACM International Conference on Advances in Soc. Networks Analysis and Min doi: 10.1109/asorn.2012.28
[4] Bandari R, Asur S and Huberman B A 2012 The Pulse of News in Social Media: Forecasting Popularity Retrieved from https://arxiv.org/pdf/1202.0332.pdf
[5] Bauer M W and Suerdem A 2016 Developing science culture indicators through text mining and online media monitoring OECD Blue Sky Forum on Science and Innovation Indicators 2016 (19–21 September 2016, Ghent, Belgium)
[6] Willaert T et al 2020 Building Social Media Observatories for Monitoring Online Opinion Dynamics Social Media+ Society. 6 (2) 2056305119898778
[7] Neresini F and Lorenzet A 2016 Can media monitoring be a proxy for public opinion about technoscientific controversies? The case of the Italian public debate on nuclear power Public Understanding of Science, 25 (2) pp 171–85
[8] Thanasopon B et al 2017 Extraction and evaluation of popular online trends: A case of Pantip.com 9th International Conference on Information Technology and Electrical Engineering (ICITEE) IEEE pp 1–5
[9] Macharia S 2020 Global Media Monitoring Project (GMMP) The International Encyclopedia of Gender, Media, and Communication pp 1–6
[10] Barile F et al 2019 A news recommender system for media monitoring IEEE/WIC/ACM International Conference on Web Intelligence pp 132–40
[11] Guo Y, Barnes S J and Jia Q 2017 Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation Tourism Management, 59 pp 467–83
[12] Curiskis S A et al 2020 An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit Information Processing & Management, 57 (2) 102034
[13] Basnyat B et al 2017 Analyzing social media texts and images to assess the impact of flash floods in cities IEEE International Conference on Smart Computing (SMARTCOMP) IEEE pp 1–6
[14] Mashechkin Igor’, Mikhail P and Dmitriy T 2013 Metody vychisleniya relevantnosti fragmentov teksta na osnove tematicheskikh modeley v zadache avtomaticheskogo annotirovaniya Vychislitel’nyye metody i programmirovaniye, 14 (1) pp 91–102
[15] Devlin J et al 2018 Bert: Pre-training of deep bidirectional transformers for language understanding arXiv (preprint arXiv:1810.04805)
[16] Parkhomenko P, Artur G and Nikita A 2017 Obzor I eksperimental’noye sravneniye metodov klasterizatsii tekstov Trudy Instituta sistemnogo programmirovaniya RAN, 29 (2) pp 161–200
[17] Vorontsov K V and Potapenko A A 2012 Regularization, robustness and sparseness of probabilistic thematic models (in Russian) Computer Res. and Modeling, 4 (4) pp 693–706
[18] Vorontsov K et al 2015 Bigartm: Open source library for regularized multimodal topic modeling of large collections International Conference on Analysis of Images, Soc. Networks and Texts (Springer Cham) pp 370–381
[19] Blei D, Andrew Y and Michael J 2003 Latent Dirichlet allocation J. of Machine Learning Res pp 993–1022
[20] Mukhamediev R I et al 2019 Multi-Criteria Spatial Decision Making Support system for Renewable Energy Development in Kazakhstan IEEE Access, 7 pp 122275-122288 doi: 10.1109/ACCESS.2019.2937627
[21] Muhamediev R I, Musabaev R R, Buldybaev T, Kuchin YA, Symagulov A, Ospanova, Yakunin K, Murzahmetov S and Sagyndyk B 2020 Ekspereimenty po ocenke sredstv massovoy informacii na osnove tematicheskoj modeli korpusa tekstov Cloud of Science, 7 (1) pp 87–104 (In Russian)
[22] Barakhnin V B et al 2019 The design of the structure of the software system for processing text document corpus Biznes-informatika, 13 (4)