data

MDPI

Data Descriptor

KazNewsDataset: Single Country Overall Digital Mass Media Publication Corpus

Kirill Yakunin [1,2,*], Maksat Kalimoldayev [1], Ravil I. Mukhamediev [1,2,3,*], Rustam Mussabayev [1], Vladimir Barakhnin [4,5,*], Yan Kuchin [1], Sanzhar Murzakhmetov [1], Timur Buldybayev [6], Ulzhan Ospanova [6], Marina Yelis [2,*], Akylbek Zhumabayev [2], Viktors Gopejenko [3,7], Zhazirakhanym Meirambekkyzy [1] and Alibek Abdurazakov [2]

1 Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan; mnk@ipic.kz (M.K.); rustam@iict.kz (R.M.); ykuchin@mail.ru (Y.K.); sanzharmrz@gmail.com (S.M.); zh.meirambekkyzy@ipic.kz (Z.M.)
2 Institute of Cybernetics and Information Technology, Satbayev University (KazNRTU), Almaty 050013, Kazakhstan; a.zhumabayev@satbayev.university (A.Z.); a.abdurazakov@satbayev.university (A.A.)
3 Department of Natural Science and Computer Technologies, ISMA University, LV-1011 Riga, Latvia; viktors.gopejenko@venta.lv
4 Federal Research Center for Information and Computational Technologies, 630090 Novosibirsk, Russia
5 Department of Information Technologies, Novosibirsk State University, 630090 Novosibirsk, Russia
6 Information-Analytical Center, Nur-Sultan 010000, Kazakhstan; Timur.Buldybayev@iac.kz (T.B.); Ulzhan.ospanova@iac.kz (U.O.)
7 International Radio Astronomy Centre, Ventspils University of Applied Sciences, LV-3601 Ventspils, Latvia
* Correspondence: Yakunin.k@mail.ru (K.Y.); ravil.muhamedyev@gmail.com (R.I.M.); bar@ict.nsc.ru (V.B.); 921126400115-D@stud.satbayev.university (M.Y.)

Citation: Yakunin, K.; Kalimoldayev, M.; Mukhamediev, R.I.; Mussabayev, R.; Barakhnin, V.; Kuchin, Y.; Murzakhmetov, S.; Buldybayev, T.; Ospanova, U.; Yelis, M.; et al. KazNewsDataset: Single Country Overall Digital Mass Media Publication Corpus. *Data* **2021**, *6*, 31. https://doi.org/10.3390/data6030031

Academic Editor: Erik Cambria

Received: 30 December 2020
Accepted: 10 March 2021
Published: 14 March 2021

Abstract: Mass media is one of the most important elements influencing the information environment of society. The mass media is not only a source of information about what is happening but is often the authority that shapes the information agenda, the boundaries, and forms of discussion on socially relevant topics. A multifaceted and, where possible, quantitative assessment of mass media performance is crucial for understanding their objectivity, tone, thematic focus and, quality. The paper presents a corpus of Kazakhstan media, which contains over 4 million publications from 36 primary sources (which has at least 500 publications). The corpus also includes more than 2 million texts of Russian media for comparative analysis of publication activity of the countries, also about 4000 sections of state policy documents. The paper briefly describes the natural language processing and multiple-criteria decision-making methods, which are the algorithmic basis of the text and mass media evaluation method, and describes the results of several research cases, such as identification of propaganda, assessment of the tone of publications, calculation of the level of socially relevant negativity, comparative analysis of publication activity in the field of renewable energy. Experiments confirm the general possibility of evaluating the socially significant news, identifying texts with propagandistic content, evaluating the sentiment of publications using the topic model of the text corpus since the area under receiver operating characteristics curve (ROC AUC) values of 0.81, 0.73 and 0.93 were achieved on abovementioned tasks. The described cases do not exhaust the possibilities of thematic, tonal, dynamic, etc., analysis of the considered corpus of texts. The corpus will be interesting to researchers considering both multiple publications and mass media analysis, including comparative analysis and identification of common patterns inherent in the media of different countries.

Dataset: https://data.mendeley.com/datasets/2vz7vtbhn2/1; https://data.mendeley.com/datasets/hwj24p9gkh/1

Dataset License: : CC-BY-SA

Keywords: natural language processing; mass-media; topic modeling; LDA; ARTM; multiple-criteria decision-making (MCDM); computer modeling; sentiment analysis; significant social news; propaganda identification

1. Summary (Required)

The mass media are a source of documents, which allows us to stay up-to-date, form our own judgments and opinions on certain events or decisions, and develop certain media consumption habits. However, this increases the possibility of spreading distorted, biased information, erroneous information, and, finally, the manipulation of information consumers.

Evaluating the impact of mass media requires rapid processing of large amounts of textual information, which can be achieved using natural language processing (NLP) and machine learning (machine learning -ML) techniques. These technologies allow users to extract information from large amounts of textual data [1,2], provide content analysis [3,4], personalized access to news [5–7], and even support its production and distribution [8,9].

Natural language processing (NLP) as a field of research includes a wide range of application areas:

- Automatic translation [10];
- Automatic summarization;
- Generating responses to user requests (question answering) [11];
- Data mining (IE) [12];
- Information retrieval [13,14];
- Sentiment analysis [15];
- Other areas are in some way related to the processing of spoken and written natural language.

NLP as a research field is changing very dynamically. Since [16], qualitatively new results have been obtained in the development of statistical language models. The large volumes of available texts in social networks and the usage of deep neural networks [17] lead to the formulation of image extraction tasks from vast amounts of unstructured information based on modern methods of distributed linguistics and so-called supervised learning.

Another modern approach is using concept-level knowledge databases, such as WordNet, ConceptNet and SenticNet [18]. This approach proves to be highly efficient and interpretable; however, it requires a knowledge-base to be developed for a certain language. A solution for this problem was proposed in [19]. In addition, a survey on multilingual sentiment analysis, including proposing solutions for scarce resource languages, was published in [20].

The key aspects that have led to impressive results in automatic natural language processing are, according to [21], advances in the development of machine-learning methods, especially deep learning [22,23], the multiplication of computing power, the availability of large amounts of linguistic data and the development of an understanding of natural language structure as applied to the social context.

However, the volumes of textual information collected are often narrowly focused [24–28]. In contrast, the Kazakh mass media corpus described below is broadly applicable. Below we describe the dataset's content, some models, and problem-solving methods using this largely unlabeled corpus.

2. Data Description (Required)

The dataset contains news publications from publicly available news websites as well as from social networks, including VK.com, YouTube, Instagram and Telegram. The corpus is presented in two forms:

The first one contains only basic meta-information about each publication, such as title, source and publication date and time and consists of Kazakhstani and Russian news.

It can be used, for example, for combative analysis of the news of the two countries or any other text-related tasks (topic-modeling, information retrieval, etc.).

The second form consists of news only from Kazakhstani sources but contains three additional columns groups:

(1) Weights of correspondence to handpicked topic groups;
(2) Weights of correspondence to 200 topics from BigARTM model;
(3) Type—either news publication or governmental program document.

It should also be noted that the dataset was not manually verified or edited. Due to technical difficulties and limitations. Hence, the dataset can contain:

- HTML and JavaScript code pieces;
- Wrong publication date and time due to format issues;
- Wrong URL;
- In rare cases, a text field can contain text from different publications.

However, such cases are not frequent and, according to verification on a small subset of data, only occur in less than 5% of publications.

2.1. Form 1 of Corpus Representation

Corpus of news from Russian and Kazakhstani news sources from 2000 to 2020 from 50 major sources, including social networks (VK.com, YouTube, Instagram and Telegram) and news websites. It includes 4,233,990 documents from Kazakhstani sources and 2,027,963 documents from Russian sources (Figure 1).

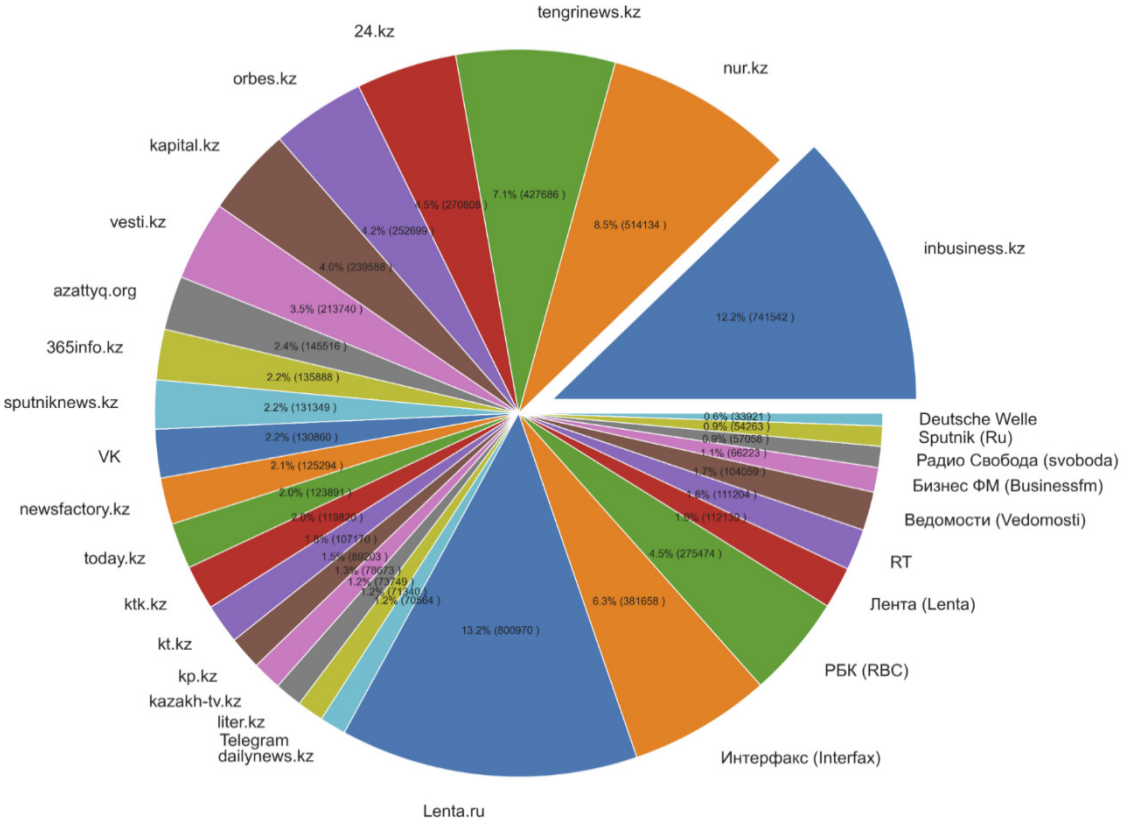

Figure 1. Major sources of the corpus.

It is available at [29]. For each document, the following fields are included:

- ID;
- Title;
- Text;

- Source;
- URL;
- Publication date and time;
- Number of views.

Tables 1 and 2 show the top sources for Kazakhstani and Russian news presented in the corpus. The full list of sources goes as follows:

Table 1. Top 20 Kazakhstani news sources by volume of publications.

Source	Number of Documents
https://inbusiness.kz/ru, accessed on 14 March 2021	741,542
https://www.nur.kz/, accessed on 14 March 2021	514,134
https://tengrinews.kz/, accessed on 14 March 2021	427,687
https://24.kz/ru/, accessed on 14 March 2021	270,809
https://forbes.kz/, accessed on 14 March 2021	252,699
https://kapital.kz/, accessed on 14 March 2021	239,588
http://vesti.kz/, accessed on 14 March 2021	213,740
https://rus.azattyq.org/, accessed on 14 March 2021	145,517
https://365info.kz/, accessed on 14 March 2021	135,889
https://sputniknews.kz/, accessed on 14 March 2021	131,350
VK	130,860
http://www.newsfactory.kz/, accessed on 14 March 2021	125,294
http://today.kz, accessed on 14 March 2021	123,892
https://www.ktk.kz/, accessed on 14 March 2021	119,820
https://www.kt.kz/, accessed on 14 March 2021	107,171
http://www.kp.kz/, accessed on 14 March 2021	89,203
https://kazakh-tv.kz/ru, accessed on 14 March 2021	78,674
https://liter.kz/, accessed on 14 March 2021	73,749
Telegram	71,340
http://www.dailynews.kz/, accessed on 14 March 2021	70,564

Table 2. Top 10 Russian news sources by volume of publications.

Source	Number of Documents
Lenta.ru	800,970
Интерфакс (Interfax)	381,659
РБК (RBC)	275,474
Лента (Lenta)	112,139
RT	111,204
Ведомости(Vedomosti)	104,059
Бизнес ФМ (Businessfm)	66,224
Радио Свобода (svoboda)	57,059
Sputnik (Ru)	54,264
Deutsche Welle	33,922

Kazakhstani sources:
https://inbusiness.kz/ru, https://www.nur.kz/, https://tengrinews.kz/, https://24.kz/ru/, https://forbes.kz/, https://kapital.kz/, http://vesti.kz/, https://sputniknews.kz/, https://rus.azattyq.org/, https://365info.kz/, VK, http://today.kz, http://www.newsfactory.kz/, https://www.ktk.kz/, http://www.kp.kz/, https://www.kt.kz/, https://kazakh-tv.kz/ru, Telegram, https://liter.kz/, http://www.dailynews.kz/, https://bnews.kz/ru (baigenews.kz), https://www.zakon.kz/, https://vlast.kz/, http://www.spik.kz/, https://rezonans.kz/, http://kz.mir24.tv/, https://kaztrk.kz/ru, egemen.kz, https://aif-kaz.kz/, https://www.interfax.kz/, Instagram, https://menshealth.kz/, http://rk-news.com/, http://infonedra.kz/, https://kz.expert/, YouTube, https://www.kazpravda.kz/, https://www.sports.kz/, http://alashinform.kz/, https://baribar.kz/,

https://korrespondent.net/, VKontakte, http://edunews.kz/, Facebook, https://adebiportal.kz/kz, http://turantv.kz/, all accessed on 14 March 2021

Russian sources:

Lenta.ru, *Интерфакс*, *РБК*, RT, *Лента*, *Ведомости*, *Бизнес ФМ*, *Радио Свобода*, Sputnik (Ru), Deutsche Welle (DW), *Настоящее время*

2.2. Form 2 of Corpus Representation

Corpus of news from Kazakhstani sources from 2018 to 2020. It is available at [30]. It includes 1,142,735 documents from news web sites and social networks with the same data as in the corpus described above with an addition of:

- Sixty-seven columns with handpicked and topic groups weights with semantic names (group economy, group politics, etc.). They were normalized to range from 0 to 1;
- Two hundred columns with topic weights were obtained through topic modeling. These columns represent a theta-matrix of the topic model;
- Type—either "news" or "governmental program".

Full list of the manually obtained topic groups: accidents, agriculture, animal-protection, astrology-magic, aviation, banking-system, cars, cars-legislation, celebrations-holidays, celebrity, cinema, congratulations, coronavirus, corruption, countryside, criminal-trials, defense, diet, ecology, economy, education, employment, fashion, food-recipes, governmental-services, health, health-advice, healthcare, heritage, history, infrastructure, innovation-digitalization, international-politics, Kyrgyzstan, language, law enforcement, legislation, migration, mining, near-east, parenthood, pets, politics, popular-science, public-utilities, railway, rallies-opposition, relationships, religion, road-accidents, Russia, science-technologies, show-business, soviet-history, sports, tourism, Ukraine, urban-improvement, urban-improvement, USA, violent-crimes, volunteers, weather, welfare, work, youth, zoo.

The corpus also includes about 4000 documents, which represent fragments of governmental development program documents. Twenty-five governmental development programs were used, including 18 programs for separate regions and big cities, two long-term state development programs and five thematic programs (digitalization, countryside development, education, healthcare and welfare programs). They were manually divided into 4000 independent fragments by experts. The reason for such preprocessing is that governmental programs a generally very lengthy and can be very thematically diverse, which complicates usage of the whole documents in topic modeling.

Comparison of representation of these governmental documents in topics is one of the approaches used to evaluate the social significance of news in [31].

There are also two additional files for that corpus:

- Topic-words.json file represents words with weights for the 200 topics obtained through topic-modeling. It is a compressed representation of a phi matrix;
- Topic-expert-labeling-sentiment.json contains expert labeling of topics sentiment. It was used to obtain results described in [31].

3. Methods (Required)

The data gathering method is web-scraping of news and media websites with open free access, as well as the scraping of social networks either by a list of social network accounts (users, groups, channels, etc.) or by a list of search queries.

The scraping algorithms were implemented in the form of Apache Airflow operators. Apache Airflow is an ETL (extract-transform-load) tool, which allows to programmatically schedule and manage tasks, monitor them and handle errors and exceptions. It was used as a core ETL-solution for the mass media monitoring system "Media Analytics" [32,33].

The scraping algorithms were implemented using the Python library Scrapy 1.7.3. For web sites with dynamic content (for example, websites based on React.JS, Angular, and other modern frontend frameworks) scrapy-splash library was used along with an official Docker image for scrapy scrapinghub/splash: 3.3.1. Scrapy was applied to simulate

running the JavaScript code inside the client's web browser to obtain the website's contents (list of news publications and news publications texts along with other meta-data).

A custom configurable Scrapy Spider was implemented, which accepts a starting URL and a list of scraping rules. This approach allows minimizing the amount of necessary software development for introducing new scraping sources since it only requires a list of scraping rules and a starting URL. It is certainly possible to create a universal parser, which would be able to scrap any source of information. However, such universal parsers tend to present a much lower quality of meta-data, and the results of scraping by such universal solutions are generally unstructured or weakly structured. That is the reason for the proposed approach, which requires a set of scraping rules for each website, but has a much higher quality and precision of meta-data, including publication date and time, a number of views, author, tags, comments, likes, shares, etc.

Scraping rules are a set of CSS-selectors for each of the accessible meta-data elements of news-publications on a given website. In principle, it is also possible to use regular expressions instead of CSS-selectors; however, CSS-selector seems to be a more fitting choice since the vast majority of HTML pages are built according to a certain CSS methodology, which makes it easy to navigate through using CSS-selectors. However, in our experience, there are some rare cases in which either an application of regular expressions or some workarounds in the Spider are required.

An example of scrap rules for Deutsche Welle news media (https://www.dw.com/ru/, accessed on 13 March 2021):

- Publication date and time-.col1.group.smallList li;
- Publication text-.intro, .longText > *p*;
- Publication title-#bodyContent h1.

Another technical issue with scraping is that news websites usually implement some measures against automatic scraping and DDoS (distributed denial of service) attacks, so using random proxy servers was implemented, as well as randomly picking user-agents to reduce chances of identification and ban. In addition, Scrapy allows configuring period between requests, concurrency and other parameters, tuning, which allows minimizing the chance of being banned by a website.

Social network scraping is a more complicated problem since social networks tend to implement some technical restrictions for scraping. In most cases, scraping is only possible through either an official API (access to which may be hard to obtain and which can be very limited) or through thorough user-actions simulation through Selenium or similar software. The second option is very costly to implement, so the first approach was used wherever possible.

Topic modeling. One of the methods that are productively applied in the field of NLP is topic analysis or topic modeling (TM). TM is a method based on the statistical characteristics of document collections, which is used in the tasks of automatic summarization, information retrieval and clustering [34]. TM transforms to the algorithm the intuitive understanding that documents in a collection form groups in which the frequency of occurrence of words or word combinations differs.

The basis of TM is the statistical model of natural language. Probabilistic TM describes documents (M) by a discrete distribution on a set of topics (T) and topics by a discrete distribution on a set of the term [35]. In other words, the TM determines which topics each document applies to and which words form each topic. Clusters of terms and phrases formed in the process of thematic modeling, in particular, allow solving the problems of synonymy and polysemy of terms [36].

To build a thematic model of the corpus of documents, a very popular latent Dirichlet allocation (LDA) [37,38] is used. LDA can be expressed by the following equality:

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

which represents the sum of mixed conditional distributions on all T set topics, where $p(w \mid t)$ is the conditional distribution of words in themes, and $p(t \mid m)$ is the conditional distribution of topics in the news. The transition from conditional distribution $p(w \mid t,\ m)$ in $p(w \mid t)$ is carried out at the expense of the hypothesis of conditional independence, according to which the appearance of words in news m on the topic t depends on the topic, but does not depend on the news m, and is common for all news. This ratio is fair, based on the assumption that there is no need to maintain the order of documents (news) in the body 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 Dirichlet's continuous multidimensional probability distribution. The purpose of the algorithm is to search for parameters φ_{wt} and θ_{tm} by maximizing the likelihood function with appropriate regularization:

$$\sum_{d \in D} \sum_{w \in D} n_{dw} ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\varphi, \theta) \rightarrow max$$

The method of maximizing the coherence value based on the UMass metric is often used to determine the optimal number of topics [39]. Some generalization of LDA is additive regularization of topic models (ARTM), implemented in the form of BigARTM library [40]. The LDA method is described in more detail in Appendix A.

The described models, as well as some methods of multiple-criteria decision-making (MCDM): AHP (analytical hierarchy process) [41], Bayesian networks [42,43], are used to classify socially significant news [31], identify propaganda [44], assess the sentiment of publications [45,46], comparative analysis of publication activity in the field of renewable energy in Russia and Kazakhstan [47]. The details of the method are described in [48,49].

Another possible application is the assessment of the dynamics of publication activity in certain topics or regarding certain persons, organizations, and events. It can be used as a numerical evaluation of the popularity of certain topics and entities, for example, in humanitarian research and for practical applications, such as public-relation department effectivity estimation (KPI), reputation management, competitors' analysis, etc.

However, it should be noted that the absolute number of publications is not a representative estimate since publication activity in online media is growing rapidly, as illustrated in Figure 2—during the last ten years, the number of publications has grown tenfold. Hence, a normalization is required, which should take into account the overall number of publications in a given period.

Figure 2. Weekly publication activity from 2010 to 2020.

Such normalization allows obtaining much more representative results. For example, Figure 3 illustrates normalized weekly publication activity on topics related to viruses and infections over the last ten years. It is obvious that publication activity on this topic was very stable over the years and almost doubled during the COVID-19 outbreak in yearly 2020.

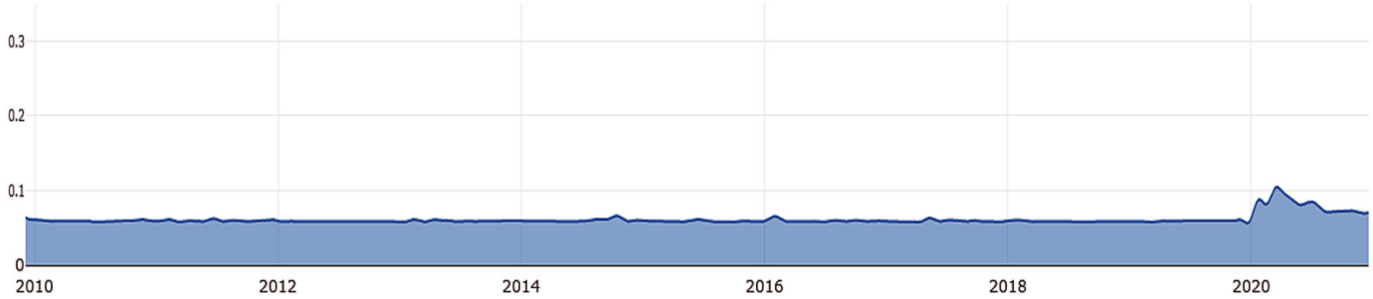

Figure 3. Normalized weekly publication activity in topics related to viruses and infections.

4. Limitations of the Study

The presented corpus has the following limitations:

- The results of model verification are based on datasets, each of which was labeled by a single expert. No thorough validation of expert's assessments was performed, only visual validation.
- The volume of the labeled subset is small compared to the volume of the corpus.

5. Conclusions

The paper described a text corpus, which contains over 4 million publications of Kazakhstani media, more than 2 million texts of Russian media and about 4000 sections of state development program documents. The corpus was used in several research cases, such as identification of propaganda, assessment of the sentiment of publications, calculation of the level of socially significant negativity, comparative analysis of publication activity in the field of renewable energy. Experiments confirm the general possibility of evaluating the social significance of news using the topic model of the text corpus since an area under receiver operating characteristics curve (ROC AUC) score of 0.81 was achieved in the classification task, which is comparable with results obtained for the same task by applying the bidirectional encoder representations from transformers (BERT) model. The proposed method of identifying texts with propagandistic content was cross-validated on a labeled subsample of 1000 news and showed high predictive power—ROC AUC 0.73. In the task of sentiment analysis, the proposed method showed a 0.93 ROC AUC score.

Despite the noted limitations, the corpus will be of interest to researchers analyzing media, including comparative analysis and identification of common patterns inherent in the media of different countries.

One of the directions of further research of the corpus is the analysis of publication activity related to individual organizations, topics and events, for example, healthcare and the COVID-19 pandemic.

Author Contributions: Conceptualization, R.I.M., K.Y. and R.M.; methodology, R.I.M. and R.M.; software, K.Y. and S.M.; validation, V.G., T.B., U.O. and Y.K.; formal analysis, R.I.M.; investigation, K.Y., V.B., M.Y., Y.K., T.B., U.O., A.Z., V.G., A.A. and S.M.; resources, M.K., R.M.; data curation, K.Y., Z.M., A.A., A.Z. and S.M.; writing—original draft preparation, R.I.M. and K.Y.; writing—review and editing, M.Y., V.B. and K.Y.; visualization, R.I.M. and K.Y.; supervision, R.I.M. and R.M.; project administration, M.K. and R.M.; funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science under the Ministry of Education and Science of the Republic of Kazakhstan, grant AP08856034.

Data Availability Statement: Data available in a publicly accessible repository. The data presented in this study are openly available in Mendeley Data at DOI:10.17632/hwj24p9gkh.1 and DOI:10.17632/2vz7vtbhn2.1.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. LDA and BigARTM Description

LDA can be expressed by the following equality:

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

representing the sum of mixed conditional distributions on all T set topics, where $p(w \mid t)$ the conditional distribution of words in themes, $p(t \mid m)$ conditional distribution of topics in the news. The transition from conditional distribution $p(w \mid t,m)$ in $p(w \mid t)$ is carried out at the expense of the hypothesis of conditional independence, according to which the appearance of words in news m on the topic t depends on the topic, but does not depend on the news m, and is common for all news. This ratio is fair, based on the assumption that there is no need to keep the order of documents (news) in the body 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 Dirichlet's continuous multidimensional probability distribution. The purpose of the algorithm is to search for parameters φ_{wt} and θ_{tm}, by maximizing the likelihood function with appropriate regularization:

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

where n_{mw} —number of occurrences of the word w, in the document m, $R(\varphi,\theta)$ —logarithmic regularizer. To determine the optimal number of thematic clusters T, the method of maximizing the coherence value calculated using UMass metrics is often used [31]:

$$U(w_i,w_j,\varepsilon) = log\frac{M(w_i,w_j)+\varepsilon}{M(w_j)}$$

where $M(\mathrm{w_i},\mathrm{w_j})$ —the number of documents containing the words $\mathrm{w_i}$ and $\mathrm{w_j}$, and $\mathrm{M(w_j)}$ —number of documents containing only the word $\mathrm{w_j}$. On the basis of this measure, the coherence value of a single thematic cluster is calculated:

$$Coh(W_k) = \sum_{(\mathrm{w_i w_j}) \in \mathrm{W}} U(\mathrm{w_i w_j},\ \varepsilon)$$

where W_k —cluster word count, ε smoothing factor is usually equal to 1.

The more documents with two words relative to documents containing only one word, the higher the coherence value of an individual topic. As a result, so many thematic clusters were chosen where the maximum average coherence value is reached:

$$Coh(W_k) = \underset{\mathrm{T}}{\mathrm{argmax}}\frac{1}{T}\sum_{k \in \mathrm{T}} Coh(W_k)$$

BigARTM

BigARTM is an open-source library for parallel construction of thematic models on large text cases, whose implementation is based on the additive regularization approach (ARTM), when the functionality of maximizing the logarithm of plausibility, restoring the original distribution of W words on documents D, is added a weighted sum of regularizers (2), by many criteria:

$$\sum_{d \in D} \sum_{w \in D} n_{dw} ln \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\varphi,\theta) \to max \quad R(\varphi,\theta) = \sum_{i=1} \tau_i R_i(\varphi,\theta)$$

n_{dw} —number of occurrences of the word w, in document d, φ_{wt} — word w distribution in topic t, θ_{td} — distribution of the topic t over the documents d. This summand $\sum_{i=1} \tau_i R_i(\varphi,\theta)$, is a weighted linear combination of regularizers, with a non-negative τ_i weights. BigARTM

offers a set of regularizers implemented on the basis of Kullback–Leibler divergence, in this case, demonstrating the entropic differences between the distributions of the initial matrix p′(w | d) and the model p′(w | d):

Smoothing regularizer, based on the assumption that the matrix columns φ *and* θ, are generated by Dirichlet distributions with hyperparameters $\beta_0\beta_t$ and $\alpha_0\alpha_t$ (identical to the implementation of the Dirichlet LDA latent placement model, where hyperparameters can only be positive).

$$R(\varphi,\theta) = \beta_0 \sum_{t\in T}\sum_{w\in W} \beta_{wt} ln\varphi_{wt} + \alpha_0 \sum_{d\in D}\sum_{w\in W} \alpha_{td} ln\theta_{td} \to max$$

In this way, we can highlight background topics, defining the vocabulary of the language, or calculate the general vocabulary in the section of each document.

Decreasing regularizer, reverse smoothing regularizer

$$(\varphi,\theta) = -\beta_0 \sum_{t\in T}\sum_{w\in W} \beta_{wt} ln\varphi_{wt} - \alpha_0 \sum_{d\in D}\sum_{w\in W} \alpha_{td} ln\theta_{td} \to max$$

aims to identify significant subject words, so-called lexical kernels, as well as subject topics in each document, zeroing out small probabilities.

Decorrective regularizer makes topics more "different". The selection of themes allows the model to get rid of small, uninformative, duplicate and dependent themes.

$$(\varphi,\theta) = -0.5 * \tau \sum_{t\in T}\sum_{s\in T/t} cov(\varphi_t\varphi_s) \to max,$$

$$cov(\varphi_t\varphi_s) = \sum_{w\in W} \varphi_{wt}$$

this regularizer is independent of matrix θ, here the estimation of differences in the discrete distributions is implemented $\varphi_{wt} = p(w|t)$ where the measure is—covariance of the current distribution of words in the topics φ_t versus the calculated distributions φ_s, where $s \in T/t$.

References

1. Korencÿic, D.; Ristov, S.; Sÿnajder, J. Document-based topic coherence measures for news media text. *Expert Syst. Appl.* **2018**, *114*, 357–373. [CrossRef]
2. Georgiadou, E.; Angelopoulos, S.; Drake, H. Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. *Int. J. Inf. Manag.* **2020**, *51*, 102048. [CrossRef]
3. Neuendorf, A. *The Content Analysis Guidebook*; Sage: Thousand Oaks, CA, USA, 2016.
4. Flaounas, I.; Ali, O.; Lansdall-Welfare, T.; De Bie, T.; Mosdell, N.; Lewis, J.; Cristianini, N. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content topics, style and gender. *Digit. Journal.* **2013**, *1*, 102–116. [CrossRef]
5. Steinberger, J.; Ebrahim, M.; Ehrmann, M.; Hurriyetoglu, A.; Kabadjov, M.; Lenkova, P.; Steinberger, R.; Tanev, H.; VGŸzquez, S.; Zavarella, V. Creating sentiment dictionaries via triangulation. *Decis. Support Syst.* **2012**, *53*, 689–694. [CrossRef]
6. Vossen, P.; Rigau, G.; Serafini, L.; Stouten, P.; Irving, F.; Van Hage, W.R. NewsReader: Recording history from daily news streams. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 26–31 May 2014; pp. 2000–2007.
7. Li, L.; Zheng, L.; Yang, F.; Li, T. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Syst. Appl.* **2014**, *41*, 3168–3177. [CrossRef]
8. Clerwall, C. Enter the robot journalist: Users' perceptions of automated content. *Journal. Pract.* **2014**, *8*, 519–531. [CrossRef]
9. Popescu, O.; Strapparava, C. Natural Language Processing meets Journalism. In Proceedings of the 2017 EMNLP Workshop, Copenhagen, Denmark, 7 September 2017; Association for Computational Linguistics: Vancouver, Canada, 2017.
10. Sreelekha, S.; Bhattacharyya, P.; Shishir, J.; Malathi, D. A Survey report on Evolution of Machine Translation. *Int. J. Control Theory Appl.* **2016**, *9*, 233–240.
11. Höffner, K.; Walter, S.; Marx, E.; Usbeck, R.; Lehmann, J.; Ngomo, A.N. Survey on challenges of Question Answering in the Semantic Web. *Semant. Web* **2017**, *8*, 895–920. [CrossRef]
12. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*; Pearson: London, UK, 2014; Volume 3.

13. Deo, A.; Gangrade, J.; Gangrade, S. A survey paper on information retrieval system. *Int. J. Adv. Res. Comput. Sci.* **2018**, *9*, 778. [CrossRef]
14. Shokin, Y.I.; Fedotov, A.M.; Barakhnin, V.B. *Problems in Finding Information*; Science: Novosibirsk, Russian, 2010; p. 220.
15. Sun, S.; Luo, C.; Chen, J. A review of natural language processing techniques for opinion mining systems. *Inf. Fusion* **2017**, *36*, 10–25. [CrossRef]
16. Manning, C.; Schutze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
17. Goldberg, Y. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420. [CrossRef]
18. Cambria, E. Affective Computing and Sentiment Analysis. In *IEEE Intelligent Systems*; IEEE: Piscataway, NJ, USA, 2016; Volume 31, pp. 102–107. [CrossRef]
19. Vilares, D.; Peng, H.; Satapathy, R.; Cambria, E. BabelSenticNet: A Commonsense Reasoning Framework for Multilingual Sentiment Analysis. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1292–1298. [CrossRef]
20. Lo, S.L.; Cambria, E.; Chiong, R.; Cornforth, D. Multilingual sentiment analysis: From formal to informal and scarce resource languages. *Artif. Intell. Rev.* **2017**, *48*, 499–527. [CrossRef]
21. Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science* **2015**, *349*, 261–266. [CrossRef]
22. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
23. Zhang, Q.; Yang, L.; Chen, Z.; Li, P. A survey on deep learning for big data. *Inf. Fusion* **2018**, *42*, 146–157. [CrossRef]
24. Coronavirus Tweets NLP—Text Classification. Available online: https://www.kaggle.com/datatattle/covid-19-nlp-text-classification (accessed on 12 March 2021).
25. Spam Text Message Classification. Available online: https://www.kaggle.com/team-ai/spam-text-message-classification (accessed on 12 March 2021).
26. Open Food Facts. Available online: https://www.kaggle.com/openfoodfacts/world-food-facts (accessed on 12 March 2021).
27. Getting Real about Fake News. Available online: https://www.kaggle.com/mrisdal/fake-news (accessed on 12 March 2021).
28. Credit Card Fraud Detection. Available online: https://www.kaggle.com/konradb/text-recognition-total-text-daset (accessed on 12 March 2021).
29. Kazakhstani and Russian News Corpus. 2020. Available online: https://data.mendeley.com/datasets/2vz7vtbhn2/1 (accessed on 12 March 2021).
30. Kazakhstani News Corpus for Social Significance Identification with Topic Modelling Results. 2020. Available online: https://data.mendeley.com/datasets/hwj24p9gkh/1 (accessed on 12 March 2021).
31. Mukhamediev, R.I.; Yakunin, K.; Mussabayev, R.; Buldybayev, T.; Kuchin, Y.; Murzakhmetov, S.; Yelis, M. Classification of Negative Information on Socially Significant Topics in Mass Media. *Symmetry* **2020**, *12*, 1945. [CrossRef]
32. Barakhnin, V.; Kozhemyakina, O.; Mukhamediev, R.; Borzilova, Y.; Yakunin, K. The design of the structure of the software system for processing text document corpus. *Bus. Inform.* **2019**, *13*, 60–62. [CrossRef]
33. Yakunin, K. Media Monitoring System. Available online: https://github.com/KindYAK/NLPMonitor (accessed on 14 September 2020).
34. Mashechkin, I.V.; Petrovskiy, M.I.; Tsarev, D.V. Methods for calculating the relevance of text fragments based on thematic models in the problem of automatic annotation. *Comput. Methods Program.* **2013**, *14*, 91–102.
35. Vorontsov, K.V.; Potapenko, A.A. Regularization, robustness and sparseness of probabilistic thematic models. *Comput. Res. Modeling* **2012**, *4*, 693–706. [CrossRef]
36. Parhomenko, P.A.; Grigorev, A.A.; Astrakhantsev, N.A. A survey and an experimental comparison of methods for text clustering: Application to scientific articles. *Proc. Inst. Syst. Program. RAS* **2017**, *29*, 161–200. [CrossRef]
37. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
38. Hamed, J.; Yongli, W.; Chi, Y.; Xia, F. Latent Dirichlet Allocation (LDA) and Topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2017**, *78*, 15169–15211.
39. Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Scotland, UK, 27–31 July 2011; pp. 262–272.
40. Vorontsov, K.; Frei, O.; Apishev, M.; Romov, P.; Dudarenko, M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. In *International Conference on Analysis of Images, Social Networks and Texts*; Springer: Cham, Switzerland, 2015; pp. 370–381.
41. Saaty, T. *Group Decision Making and the AHP*; Springer: New York, NY, USA, 1989.
42. Mohammad, A.; Ehsan, A.; Rouzbeh, A.; Vikram, G.; Irene, P. Developing a Novel Risk-based Methodology for Multi-Criteria Decision Making in Marine Renewable Energy Applications. *Renew. Energy* **2017**, *102*, 341–348. [CrossRef]
43. Mukhamediev, R.I.; Mustakayev, R.; Yakunin, K.; Kiseleva, S.; Gopejenko, V. Multi-Criteria Spatial Decision Making Support System for Renewable Energy Development in Kazakhstan. *IEEE Access* **2019**, *7*, 122275–122288. [CrossRef]
44. Yakunin, K.; Ionescu, M.; Murzakhmetov, S.; Mussabayev, R.; Filatova, O.; Mukhamediev, R. Propaganda Identification Using Topic Modelling. *Procedia Comput. Sci.* **2020**, *178*, 205–212. [CrossRef]
45. Mukhamediev, R.; Musabayev, R.; Buldybaev, T.; Kuchin, Y.; Symagulov, A.; Ospanova; Yakunin, K.; Murzakhmetov, S.; Sagyndyk, B. Media assessment experiments based on a thematic corpus model. *Cloud Sci.* **2020**, *7*, 87–104.

46. Yakunin, K.; Mukhamediev, R.; Kuchin, Y.; Musabayev, R.; Buldybayev, T.; Murzakhmetov, S. Classification of negative publication in mass media using topic modeling. *J. Phys. Conf. Ser* **2020**. in print.
47. Yakunin, K.; Musabaev, R.; Yelis, M.; Mukhamediev, R. The topic of energy in news publications. In Proceedings of the All-Russian Scientific Conference and the xiii Youth School with International Participation, Moscow, Russian, 24–25 November 2020; pp. 451–456.
48. Musabaev, R.; Muhamedyev, R.; Kuchin, Y.; Symagulov, A.; Yakunin, K. On a method of multimodal media ranking using corpus based topic modelling. *Inf. Technol. Manag. Soc.* **2019**, *4*, 5.
49. Yakunin, K.; Mukhamediev, R.; Mussabayev, R.; Buldybayev, T.; Kuchin, Y.; Murzakhmetov, S.; Rassul, Y.; Ospanova, U. Mass Media Evaluation Using Topic Modelling. In *International Conference on Digital Transformation and Global Society*; Springer: Cham, Switzerland, 2020; pp. 130–135.