Applying text mining methods to extracting information from news articles

Tsvetanka Georgieva-Trifonova¹, Miroslav Dechev¹
¹“St. Cyril and St. Methodius” University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria

e-mail: cv.georgieva@live.uni-vt.bg

Abstract. In the present paper, the text mining methods from the point of view of their application for extracting information from news articles, are discussed. For this purpose, research questions are formulated and opportunities to assist in obtaining their respective answers are presented, as they are illustrated by the application of text mining methods on a created dataset with articles in Bulgarian. The extracted information could be useful for research of the topics covered in the articles, exploration of customer interests, etc.

1. Introduction
The growing size of news content collections, made available by digital technologies, allows and requires scientific research, challenging researchers to analyze text data [1, 2]. As a result, it is becoming increasingly important to propose a variety of applications and tools that perform methods for automated and semi-automated analysis of text documents.

On the other hand, text mining is a scientific field that provides methods for analyzing and processing unstructured data, which form nearly 95% of big data [3]. The application of text mining methods includes [4]:

• Collecting unstructured data from multiple data sources such as plain text, web pages, pdf files, emails and blogs, etc.;
• Discovering and removing data anomalies by performing operations for pre-processing and cleaning;
• Transforming the necessary information extracted from unstructured data into structured formats;
• Applying a specific text mining algorithm for extracting additional information;
• Analyzing and interpreting the discovered models in the data;
• Storing the found valuable information in a database to provide an opportunity to analyze trends and improve the decision-making process.

Text mining methods allow extracting information from unstructured data to obtain:

• Data description and summarization;
For this purpose, clustering a document collections is applied; feature selection; topic modeling, i.e. identifying common themes in large collections of documents.
• Data prediction.
In this case, an association analysis is performed; classification of text documents, i.e. associating documents with predefined categories.

The present study summarizes the problems for which text mining algorithms are applied to extract information from news articles. The benefits of the obtained additional information are justified and illustrated with specific examples from a dataset created for this purpose.
The rest of the paper is organized as follows. Section 2 reviews existing research on text mining methods and their application for extracting information from news articles. Section 3 describes a dataset with news articles in Bulgarian; research questions are formulated and discussed, for the solution of which text mining algorithms are used.

2. Related work about text mining methods and their application for extracting information from news articles

There is a tendency to increase the interest in research about information extraction from the text of news and other articles based on statistical models present in the texts under consideration [5, 6]. [5] summarizes the challenges that libraries and archives pose with this type of research, such as access and licensing, copyright, technology and software support, storage and access issues, and the need for references, instructions, and training.

Researchers consider the application of text mining in terms of various aspects:

A. Subject area and topic;

[7] illustrates the usage of text mining for news and blogs, highlighting the specifics that arise from the domain, including data, tasks, users and use cases.

In [8], an exploratory analysis is performed based on the application of text mining for news articles on "water" and "society". The online archives of the newspapers the Japan News and the International New York Times are used for this purpose. Clustering is applied to divide the collection of documents into mutually exclusive groups based on the presence of similar topics.

[9] examines trends in consumer policy studies, extracting the titles and keywords of articles in the journal “Consumer Policy and Education Review” and subjecting them to clustering and association analysis.

The extraction of information from scientific papers found from known scientific databases in the search for "Mobile Learning in higher education" is presented in [10]. The most frequent words in the collected papers are identified; the joint frequent occurrences of words are determined by discovering association rules; the groups of similar papers are established by clustering.

B. Methods.

In [11], the content analysis and the text mining are compared. The results obtained from the application of both qualitative data analysis with Atlas.ti and text mining to the same set of company reports are provided in order to identify trends in the sustainability disclosure.

Commonly used text mining methods are:

• Topic modeling;

Topic modeling is performed in research that applies it to various types of content such as newspapers, scientific publications, literary texts, social media. In [12], a study is made of the successful application of topic modeling for large collections of text documents to support researchers in making insights and understanding fiction, non-fiction, scientific publications, political texts.

[13] discusses the Latent Dirichlet Allocation (LDA) topic modeling method as a useful content analysis tool designed to automatically organize large document archives based on affected topics derived from co-occurrence of words. This paper explicates the essence of LDA method; peculiarities in its application; setting the different parameters by the researcher and their influence on the results; meaningful interpretation of the obtained results.

Another in-depth study of LDA, in particular some of its strengths and weaknesses in its application in media and communication, is conducted in [14].

• Association analysis;

The use of association analysis is discussed in [15]. The Apriori algorithm for identifying association rules is applied for examination of the presence of trends and patterns in features describing news articles.

• Clustering.

In [16], clustering of text data from Facebook is applied in order to analyze them and extract useful information.
In the present paper, the application of different text mining methods in terms of extracting information from news articles is examined. The additional information obtained in this way is summarized by defining the research questions, the possibilities for solving which are discussed and illustrated with a created dataset.

3. Exploring the possibilities of text mining to extract information from news articles

For the purposes of the present study, a dataset with news articles in Bulgarian is constructed.

3.1. Dataset description

A dataset News24chasa is implemented by scraping news articles with ParseHub (www.parsehub.com). In particular, a keyword searches are conducted in the online version of the 24 Chasa newspaper (www.24chasa.bg). The keywords are science and education (наука и образование); health (здраве); sportsmen (спортисти); economics (икономика); art (изкуство); politicians (политици); crimes (престъпления). Data for articles from the first five pages with the results of each search are extracted. Thus, data are collected for a total of 1190 articles found on 15 May 2020, the earliest being from 15 June 2017. Duplicate articles found as a result of more than one search by different keywords are removed, after which the number of other different articles is 1144.

The resulting dataset includes the title of the news item; URL where it can be found; number of readings; category; date; author; full text of the article. An attribute is added containing the keywords by which the relevant article is found.

Like most similar sites, the 24 Chasa news site associates each news item with predefined categories. The collection of articles is done by keyword search, which gives more opportunities for the future usage of the created dataset, such as research related to information retrieval, news recommendation.

Preliminary processing of the collected texts is performed, which includes tokenization, removal of stop words, stemming. Table 1 contains information about the News24chasa dataset, summarized after the pre-processing. The included summaries refer to the complete dataset and to parts of it obtained according to the keywords, by which the relevant news is found.

Table 1. Summary information about the dataset News24chasa.

| Characteristic      | Number of documents (news) | Number of words | Number of unique words | Average number of words by documents | Average number of unique words by documents |
|---------------------|---------------------------|-----------------|------------------------|--------------------------------------|---------------------------------------------|
| The full dataset News24chasa | 1144                       | 381908          | 39522                  | 333.84                               | 217.75                                     |
| Science and education | 169                        | 97135           | 11996                  | 574.76                               | 300.70                                     |
| Health              | 170                        | 44340           | 10528                  | 260.82                               | 182.75                                     |
| Sportsmen           | 170                        | 51062           | 11859                  | 300.36                               | 210.99                                     |
| Economics           | 170                        | 49573           | 9689                   | 291.61                               | 199.75                                     |
| Art                 | 169                        | 68514           | 15711                  | 405.41                               | 273.26                                     |
| Politicians         | 170                        | 55696           | 12134                  | 327.62                               | 230.58                                     |
| Crimes              | 170                        | 38570           | 8636                   | 226.88                               | 156.14                                     |

3.2. Extraction of information from news articles through text mining

The following subsections discuss the application of text mining methods in terms of extracting information from news articles. The execution of the selected algorithms for text mining is performed with RapidMiner (https://rapidminer.com). Research questions are formulated, the answers received are discussed. The information extracted in this way is interpreted and illustrated using the dataset News24chasa described in subsection 3.1.
3.2.1. Finding frequent words, bigrams, trigrams of words (i.e. 2 or 3 consecutive words)
Q1: Which words and phrases are frequently used in articles found with different keywords?

A significant problem in text mining algorithms is how to make a numerical representation of a text document based on the words of which it consists. For this purpose, various measures are applied to determine the importance of a word, one of which is its frequency of occurrence in a document and in the collection of documents.

The most frequent words in the documents of the dataset News24chasa, found by keywords "health" (fig. 1), are health (здраве; it is contained in 99.41% of the selected documents, relative frequency 0.65%), coronavirus (коронавирус; 55.88% of the selected documents), measures (мерки; 40.00%), covid (45.88%), pandemic (пандемия; 38.24%); the most frequent word combinations (bigrams of words) are public_health (обществено_здраве; in 42.35% of the selected documents), emergency_position (извънредно_положение; 28.82%), new_coronavirus (нов_коронавирус; 21.18%), pandemic_covid (пандемия_covid; 11.76%).

![Figure 1](image_url)

Figure 1. Word cloud of the most frequent words in documents of dataset News24chasa found by keywords "health".

3.2.2. Finding frequent itemsets and association rules
Q2: Which words are most frequent occurred together in articles found by different keywords?

The methods for association rules mining can be applied to the text analysis, as a result of which interesting associations between the words in the documents are discovered. The association rules of text documents provide important information about the interrelationships between them, which are useful in identifying models, themes, context [17], feature selection [18], text classification [19].

The FP-Growth algorithm [20] is applied on the dataset News24chasa, which uses a tree of frequent patterns. One of the frequent itemsets for document found by searching by the keywords "science and education" is \{education (образование), science (наука), project (проект), growth (растеж), program (програма), intelligent (интелигентен)\} with the value of the parameter support 0.609467, consequently 60.9467% of the selected documents in the dataset contain all these words. From this frequent itemset, the association rule \{science (наука), education (образование)\} -> \{project (проект), growth (растеж), program (програма), intelligent (програма)\} is validated with a value of the confidence parameter 0.613095, which means that 61.3095% of the selected documents containing \{science (наука), education (образование)\}, also include the words \{project (проект), growth (растеж), program (програма), intelligent (програма)\}.

3.2.3. Clustering the entire resulting dataset
Q3: Which articles form groups of similar ones? Which words describe the articles in the same cluster? Is there a relationship between the found clusters and the keywords by which the articles are found?

Clustering is used to find groups (clusters) of documents with similar content, i.e. the documents in one group should be similar to each other and different from the documents from the other groups.
Cluster identification is performed based on selected data attributes and similarity measures. The K-means algorithm proposed in [21] is applicable to text document clustering.

For the specific dataset, the implementation of the K-means algorithm proposed in [22] is applied, which is faster especially for datasets with many attributes, as is the case with text data. For the vector representation of the documents the term frequency–inverse document frequency (tf-idf) is used; the words and bigrams of words are extracted. When setting the number of clusters K = 2 and applying to the news found by keywords "science and education", we obtain that the words and bigrams of words that are closest to the centroid of:

- cluster_0 are construction (изграждане), center (център), period (период), high (висок), economic (икономически), plan (план), organization (организация), scientific (научен);
- cluster_1 are minister_education (министър_образование), education_science (образование_наука), children (дета), science_program (програма_наука), interest (интерес), follow (следва), expect (очаква), school (училище).

They could be considered as a summary of the two most discussed topics in these articles.

![Figure 2. Presence of the word "art" in the documents from the found clusters.](image)

When the clusters are 7 and the algorithm is executed for all news in the dataset News24chasa, we can examine whether there is a relationship between the found clusters and the keywords by which the articles are found. Figure 2 shows that the word "art" ("изкуство") predominates in cluster_0, in figure 3 the bigram "science_education" ("наука_образование") stands out in cluster_6.
3.2.4. Applying a topic modeling method to articles found with different keywords;

Q4: Which words contained in the articles describe the topics covered in them?

Topic modeling is an unsupervised machine learning approach. This means that it can find patterns and group similar data without the need to pre-define topics (categories, labels) or train data. For a given set of documents, the aim is to identify the topics that cover these documents and to group them according to the topics found. The main topic is extracted, presented by a set of words that appear in the relevant documents.

The formulated question Q4 can be asked in regard to all articles or a selected part of them, for example articles found in a search by a certain keyword; articles from the same category; articles by the same author, etc. Two of the seven topics found in all News24chasa news articles through LDA algorithm [23] are described by the following words contained in the articles:

- **topic 1**: China (Китай), pandemic (пандемията), coronavirus (коронавирус), USA (САЩ), measures (мерки), epidemic (епидемията), covid, crisis (кризата), BTA (БТА);
- **topic 2**: education (образование), Bulgaria (България), technology (технологии), schools (училища), training (обучение), science (наука), university (университет), project (проекта), children (деца).

3.2.5. Feature selection

Q5: What are the most important words describing articles in regard to a selected attribute (e.g. keywords, category and author)?

The feature selection consists in selecting a subset of the words appearing in the training dataset and using only this subset both for training the model and for applying the respective algorithm for data mining.

The Relief feature selection algorithm, which is formulated in [24], is applied to the dataset News24chasa. This method extracts instance randomly from the dataset and updates the relevance of each feature, finding for the value \( x_{ij} \) of a feature \( t_i \) for a selected examples \( d_j \) (\( j=1, \ldots, n \)) its difference with that of the closest example of the same category (nearHit) and the closest example of the other category (nearMiss).
\[ RW_i = RW_i - (x_{ij} - \text{nearHit})^2 + (x_{ij} - \text{nearMiss})^2 \] (1)

When examples are text documents represented by the vector space model, \( x_{ij} \) is the weight of the word \( t_i \) for the document \( d_j \). The updating of the weight \( RW_i \) of \( t_i \) is performed according to equality (1), therefore the feature weight decreases if the considered example differs in this feature from the near examples in the same category more than from the near examples in the other category, and increases in the opposite case. The original Relief algorithm has some limitations, such as it can only be applied to two categories. A variant applicable to more categories is developed in [25]. Subsequently, Relief-based algorithms, overcoming deficiencies and limitations, are proposed. They are summarized and analysed in detail in [26].

Figure 4 shows some of the most important words found, describing news articles in regard to all keywords by which they are found.

![Figure 4. Most important words describing the news articles in regard to all keywords by which they are found when applying the Relief algorithm.](image)

3.2.6. Data classification

Q6: Is it possible to build a classification model that allows prediction with acceptable accuracy?

The classification of text documents can be applied for different defined categories, such as the keywords by which the article are found; category attribute; author. It is necessary to create a classification model that describes the articles using the words contained in them and allows prediction. The constructed model is used to recommend news, i.e. make a user recommendation for a news article that is likely to match a previous keyword search.

The decision tree text classifier [27] is based on a tree structure in which the inner vertices are labelled by terms, their branches are obtained by checking the word weights in the test dataset (set of document), the leaf vertices are the categories. For this purpose, measures for selection of an attribute for splitting are applied. The best results for the considered dataset News24chasa are obtained with gini index as a measure for defining the criterion for attribute selecting for splitting. When the category according to which the news articles are associated is the attribute with the search keywords, the accuracy of the classification is 93.71%, Macro F-measure is 93.65%.

Because deep learning approaches are among the most important and highly used in NLP (Natural Language Processing), the algorithm H2O's Deep Learning is applied [28]. H2O's Deep Learning is based on a multi-layer feedforward artificial neural network, which is trained using a back-propagation algorithm. Table 2 contains the values of measures to assess the validity of the classification – the accuracy and Macro F-measure for the K-NN (K-Nearest Neighbors) classifier, Decision tree (for two different measures to select a splitting attribute), H2O's deep learning.
Table 2. Accuracy and Macro F-measure for different classifiers.

| Classifier                | Accuracy | Macro F-measure |
|---------------------------|----------|-----------------|
| K-NN                      | 65.99%   | 65.15%          |
| Decision tree (gain ratio)| 63.55%   | 58.69%          |
| Decision tree (gini index)| 93.71%   | 93.65%          |
| H2O's deep learning       | 65.21%   | 65.04%          |

The values of these measures can be improved if a feature selection algorithm (such as that presented in subsection 3.2.5) is applied in advance. The resulting dataset would be useful in experiments designed to improve the text classification performance.

After the made study, we have not found results from the assessment of text classification applied to a dataset with news articles in Bulgarian. As a similar result, we provide the values for accuracy and Macro F-measure of the K-NN algorithm, executed for the dataset Reuters-21578, which consists of news from Reuters newswire for 1987. In [29], the K-NN algorithm is applied to a subset of the Reuters-21578 dataset containing documents of 92 category, and a Macro F-measure value of 50.89% is obtained. The result of the accuracy of the same algorithm for Reuters-21578 is presented in [30] - 66.5258%.

4. Conclusion

The development of web technologies leads to a huge interest in the processing and analysis of text documents in order to find useful additional information. Usually the approaches for text mining refer to the implementation of the process of extracting valuable information from unstructured text.

In the present paper, the possibilities of text mining to extract information from news articles are explored. Text mining algorithms are applied on a created dataset in accordance with pre-formulated research questions. In particular, algorithms for association analysis, clustering, topic modeling, feature selection, classification are implemented.

In conclusion, we can summarize that the identification of hidden information in news articles is a challenge that can be supported by implementing appropriate text mining algorithms.

References

[1] Macnamara J 2005 Media Content Analysis: Its Uses, Benefits and Best Practice Methodology Asia Pacific Public Relations Journal 6 1 pp 1–34
[2] Karlsson M and Sjøvaag H 2016 Content Analysis and Online News Digital Journalism 4 1 pp 177-192
[3] Gandomi A and Haider M 2015 Beyond the hype: Big data concepts, methods, and analytics International Journal of Information Management 35 2 pp 137-144
[4] Dang S and Ahmad P H 2014 Text Mining: Techniques and its Application IJETI International Journal of Engineering & Technology Innovations 1 4 pp 22-25
[5] Cheney D 2013 Text mining newspapers and news content: new trends and research methodologies IFLA World Library and Information Congress p 5
[6] Sohrabi B R, Vanani I and Namavar M 2019 Investigation of Trends and Analysis of Hidden New Patterns in Prominent News Agencies of Iran Using Data Mining and Text Mining Algorithms Webology 16 1 pp 114-137
[7] Berendt B 2015 Text mining for news and blogs analysis Encyclopedia of Machine Learning and Data Mining Ed C Sammut and G I Webb (Berlin: Springer)
[8] Hori S 2015 An exploratory analysis of the text mining of news articles about “water and society WIT Transactions on The Built Environment 168 pp 501-508
[9] Kim M J, Ohk K and Moon C S 2017 Trend Analysis by Using Text Mining of Journal Articles Regarding Consumer Policy New Physics: Sae Mulli 67 5 pp 555-561
[10] Salloum S A, Al-Emran M, Monem A A and Shaalan K 2018 Using Text Mining Techniques for Extracting Information from Research Articles, Intelligent Natural Language Processing;
Trends and Applications Studies in Computational Intelligence 740 (Berlin: Springer)

[11] Aureli S 2017 A comparison of content analysis usage and text mining in CSR corporate
disclosure The International Journal of Digital Accounting Research 17 pp 1-32

[12] Boyd-Graber J 2017 Applications of Topic Models Foundations and Trends in Information
Retrieval 11 2-3 pp 1-154

[13] Jacobi C, Atteveldt W and Welbers K 2015 Quantitative analysis of large amounts of journalistic
texts using topic modelling Digital Journalism 4 1 pp 89-106

[14] Puschmann C and Scheffler T 2016 Topic modeling for media and communication research: A
short primer HIIG Discussion Paper Series p 20

[15] Soni N and Ghosh S M 2016 Online News Trend Discovery Using Apriori Algorithm
International Journal for Research in Applied Science & Engineering Technology 4 6 pp 199-203

[16] Salloum S A and Al-Emran M, Shaalan K 2017 Mining Text in News Channels: A Case Study
from Facebook International Journal of Information Technology and Language Studies 1 pp 1-9

[17] Lopes A A, Pinho R, Paulovich F V and Minghim R 2007 Visual text mining using association
rules Computers & Graphics 31 3 pp 316–326

[18] Do T D, Hui S C and Fong A 2006 Associative Feature Selection for Text Mining International
Journal of Information Technology 12 4 pp 59-68

[19] Haralambous Y and Lenca P 2014 Text classification using association rules, dependency pruning
and hypernymization In Proceedings of the 1st International Conference on Interactions
between Data Mining and Natural Language Processing 1202 pp 65–80

[20] Han J and Pei J 2000 Mining frequent patterns by pattern-growth: methodology and implications
ACM SIGKDD Explorations Newsletter 2 2 pp 14-20

[21] MacQueen J B 1967 Some methods for classification and analysis of multivariate observations
Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability Ed L M Le
Cam and J Neyman 1 pp 281–297

[22] Elkan C 2003 Using the Triangle Inequality to Accelerate k-Means Proceedings of the Twentieth
International Conference on Machine Learning pp 147–153

[23] Yao L and Mimno D 2009 Andrew McCallum, Efficient Methods for Topic Model Inference on
Streaming Document Collections Proceedings of the 15th ACM SIGKDD international
conference on Knowledge discovery and data mining pp 937–946

[24] Kira K and Rendell L 1992 The Feature Selection Problem: Traditional Methods and a New
Algorithm Proceedings of the tenth national conference on Artificial intelligence pp 129-134

[25] Kononenko I 1994 Estimating attributes: Analysis and extensions of RELIEF Machine Learning:
ECML-94. Lecture Notes in Computer Science ed F Bergadano and L De Raedt 784 (Berlin:
Springer) pp 171–182

[26] Urbanowicz R J, Meeker M, La Cava W, Olson R S and Moore J H 2018 Relief-based feature
selection: Introduction and review Journal of Biomedical Informatics 85 pp 189-203

[27] Mitchell T M 1996 Machine Learning (New York: McGraw Hill)

[28] Candel A and Parmar V 2015 Deep Learning with H2O (H2O.ai, Inc.)

[29] Tan C 2006 An effective refinement strategy for KNN text classifier Expert Systems with
Applications 30 pp 290–298

[30] Pradhan L, Tanaja N A, Dixit C and Suhag M 2017 Comparison of Text Classifiers on News
Articles International Journal of Engineering and Technology 4 3 pp 2513-2517