МРНТИ 16.31.21
УДК 004.912

https://doi.org/10.51889/2959-5894.2023.81.1.020

IDENTIFYING AND ANALYZING FEATURES FOR THE CLASSIFICATION OF NEWS

Ualiyeva I.M. [1*], *Mussabayev R.R.* [2]

[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[2]*The Institute of Information and Computational Technologies, Almaty, Kazakhstan*
[*]*e-mail: i.ualiyeva@mail.ru*

Abstract

The number of documents, including online news that requires a deeper understanding and analysis grows every year. Machine Learning algorithms help us to classify texts accurately. However, finding suitable structures and techniques for text, including feature extraction, is difficult for researchers. This paper addresses the task of identifying and analyzing features to distinguish different genres of texts. We studied the main characteristics of each genre of news text like news, articles, interviews, and blogs to obtain more informative features. We have built our data set by collecting texts from open-access official information portals. Analysis of our data set and features that look at structural complexity, detail, and imaginative details in a text are helpful to distinguish our dataset. In particular, we use complexity (lexical diversity, lexical density, punctuation, average sentence length, number of personal pronouns, readability index), detail features (number of proper nouns in the text, numbers, month-related words), imaginative features (PoS tags, words-quantifiers, plural nouns) features. Our results suggest that our features provide effective representation to distinguish news texts from articles, blogs/opinions, and interviews with high accuracy.

Keywords: Text Categorization, Text Mining, Feature Selection, Text Classification, Online News Classification.

Аңдатпа

И.М. Уалиева[1], *Р.Р. Мусабаев*[2]
[1]*әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан*
[2]*Ақпараттық және есептеуіш технологиялар институты, Алматы қ., Қазақстан*

ОНЛАЙН ЖАҢАЛЫҚТАРЫН ЖІКТЕУ ЕРЕКШЕЛІКТЕРІН АНЫҚТАУ ЖӘНЕ ТАЛДАУ

Жаңартылған ақпарат көлемінің экспоненциалды өсуі ақпаратты іздеу міндетін қиындатады. Машиналық оқыту алгоритмдері мәтіндерді жіктеу арқылы іздеу кеңістігін автоматты түрде азайтуға көмектеседі. Бұл жұмыста жаңалық мәтіндерін (жаңалықтар, мақалалар, сұхбаттар және блогтар) жіктеу белгілерін анықтау, талдау және таңдау мәселесі қарастырылады. Ақпараттық белгілерін алу үшін біз жаңалықтар мәтіндерінің әрбір жанрының негізгі сипаттамаларын анықтадық. Біз ашық қолжетімділікпен ресми ақпараттық порталдардан алынған жаңалықтар корпусын жасадық және мәтіннің құрылымдық күрделілігін, егжей-тегжейлілігін және бейнелілігін қарастыратын белгілерді анықтадық. Атап айтқанда, біз күрделілік сипаттамаларын (лексикалық әртүрлілік, лексикалық тығыздық, тыныс белгілері, сөйлемнің орташа ұзақтығы, тұлғалық есімдіктердің саны, оқылу көрсеткіші), егжей-тегжейлі сипаттамалар (жалпы есімдер, сандар, айларға байланысты сөздер және т.б. саны), бейнелеу сипаттамаларын (PoS тегтері, квантор сөздері, көпше түрдегі зат есімдер) қолданамыз. Нәтижелер осы белгілердің үйлесімі жаңалықтар мәтіндерін жіктеудің жоғары дәлдігін қамтамасыз ететіндігін көрсетеді.

Түйін сөздер: онлайн жаңалықтарды жіктеу, мәтінді өңдеу, мәтінді жіктеу, мұмкіндіктерді таңдау.

Аннотация

И.М. Уалиева[1], *Р.Р. Мусабаев*[2]
[1]*Казахский Национальный Университет имени аль-Фараби, г.Алматы, Казахстан*
[2]*Институт информационных и вычислительных технологий, г.Алматы, Казахстан*

ИДЕНТИФИКАЦИЯ И АНАЛИЗ ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ ОНЛАЙН НОВОСТЕЙ

Экспоненциальный рост количества актуальной информации затрудняет задачу информационного поиска. Алгоритмы машинного обучения помогают нам автоматически снижать пространство поиска путем классификации текстов. В данной работе рассматривается задача выявления, анализа и отбора признаков для классификации новостных текстов (новости, статьи, интервью и блоги). Для получения наиболее информативных признаков мы выявили основные характеристики каждого жанра новостных текстов. Мы создали корпус новостей, взятых из официальных информационных порталов с открытым доступом, и выявили признаки, которые рассматривают структурную сложность, детализацию и образность текста. В частности, мы используем характеристики сложности (лексическое разнообразие, лексическая плотность, пунктуация, средняя длина предложения, количество личных местоимений, индекс читабельности), характеристики детализации (количество имен собственных, цифр, слов, связанных с датами и пр.), характеристики образности (PoS-теги, слова-квантификаторы, существительные во множественном числе). Результаты показывают, что совокупность этих признаков обеспечивает высокую точность классификации новостных текстов.

Ключевые слова: онлайн-новости, исследование текстов, классификация текстов, отбор признаков.

Introduction

The need for identifying and interpreting possible differences in genres of texts has increased nowadays because the number of text documents grows every day. Many researchers are now interested in developing methods to improve classification and applications that leverage text classification methods. Most text classification systems may be deconstructed as four stages: feature selection, dimensions reduction, classifier selection, and evaluations.

In this work, we have created a feature set based on past work in fake news detection, genre recognition, journalistic profile prediction, framing bias detection, and summarization. Using this feature set, we have built a model to distinguish different genres: news, articles, opinions, and interviews.

The news is information-dense text, because report important factual information in a direct, succinct manner. While the essence and perception of texts like articles, interviews, and opinion texts are more individual and personified. Articles and opinions are more subjective and represent the author's opinion while an interview represents an invited guest's opinion. The style of articles, interview, and blogs are more informal. These genres are more complex than news. But the news contains more details like dates or numbers.

So, we use the classification algorithm Random Forest and clusterization algorithms k-means++ in order to choose the best algorithm to recognize multiple genres. To make documents of different lengths comparable, each feature vector is normalized by the Max-Min Scaling method. To analyze the importance of each feature we use both algorithms.

The rest of the paper is organized as follows: firstly, we review some existing methods, for text classification by genre. The next section addresses the main differences between the four styles. In the next section, we present our approach for extracting the features to build the model. Next, we describe the classification and clusterization algorithm that we used to train our model. The last section addresses the result and the evaluation methods for our model. Finally, we conclude the paper and discuss the future work.

Related Work

Feature selection for text genre identification is studied by Kessler et al. [1], who investigate generic cues, the 'observable' properties of a text that are associated with facets. They called facets three types of text: Brow (texts that required the intellectual background of the target audience), Narrative (text is written in a narrative mode), and Genre that includes reportage, editorial, SciTech, legal, nonfiction, and fiction.

Fred Morstatter et al. [2] studied how a set of features can handle framing bias in online news. They defined multi-lingual feature groups such as unigrams, bigrams, Part-Of-Speech unigrams, and bigrams, and quotes that can be automatically extracted in any language. They also considered sentence complexity and named entities as features of framing bias. They found that simple linguistic features perform best in this classification task and that n-grams can give reasonable predictions in finding frames in text.

Annie Louis and Ani Nenkova [3] used several classes of features that capture lexical and syntactic information, as well as word specificity and polarity to classify the distribution of general and specific sentences of news articles for a task of (abstractive) summarization.

Momchil Hardalov et al. [4] studied the problem of finding fake online news. They use linguistic (n-gram), credibility-related (capitalization, punctuation, pronoun use, sentiment polarity), and semantic (embeddings and DBPedia data) features for automatically distinguishing credible news from fake news. Edward Dearden and Alistair Baron [5] created a feature set for tasks of deception detection, humor recognition, and satire detection tasks. Yatsko [6] described an experimental method for automatic text genre recognition based on forty-five statistical, lexical, syntactic, positional, and discursive parameters. They analyzed parameters that are the most significant for scientific, newspaper, and artistic texts. Adaptive summarization algorithms have been developed based on these parameters. Predicting journalistic profiles is another task when features-based approach is used. Some researchers use the frequency of POS tags [7] to classify user profiles. Others use an average number of words per sentence, the average number of letters in a word, and punctuation [8]. Daniela Gîfu and Dan Cristea [9] established a number of syntactic, lexical-semantic, and pragmalinguistic features such as personal pronouns, to predict journalistic profile.

Data Set

The data for our experiments comes from the government informational portal *tengrinews.kz*. We have selected a corpus of texts published in the recent year. Data did not require manual marks, because they were marked automatically by journalists of the portal such as *news*, *interview*, *articles*, and *blogs/opinions*.

The news are information-dense texts, because report important factual information in direct, succinct manner. While the essence and perception of texts like articles, interviews, and opinion texts are more individual and personified. Articles and opinions are more subjective and represent the author's opinion while an interview represents an invited guest's opinion.

In the articles, the author analyzes social situations, processes, and phenomena, and reasonably expresses his point of view, based on a deep analysis of facts. Articles are characterized by a clear social orientation. For example, the authors of the portal discuss socially significant issues from the organization of the workspace, the rules of conduct during an earthquake, poaching, and ending with the brain drain.

In an interview, a journalist invites a socially significant person to discuss current issues of society in a conversation with him. For example, in an interview with a political scientist, acute problems of society and how to solve them can be discussed. What are the expectations of the population and how far can the next project be successfully implemented?

In blogs, small author's stories, it is told about events, traditions, and memoirs of the author. In blogs, authors express a personal point of view, and the style of expression of the blogger is more imaginative.

Finally, our balanced corpus of news, interview, articles, and opinion texts contains 817 articles, where 219 of them are articles, blogs/opinions – 158, interviews – 220, and news – 220.

Feature Set

We used several properties of the different types of texts to encode texts as vectors of features. We hypothesized that these features may differentiate these types of texts.

To measure various dimensions of lexical richness we calculated lexical density [10], lexical diversity [10], and readability [11]. Lexical diversity is a measure of how many different words are used in a text, while lexical density provides a measure of the proportion of lexical items (i.e. nouns, verbs, adjectives, and some adverbs) in the text.

The traditional measure of lexical diversity is the type-token ratio (TTR) calculated as the total number of unique words divided by the total number of words (tokens).

However, it is not a good fit in our case when different types of texts with different sizes are compared because the values are inversely proportional to the text size. Following [12] [13], and [14], we also use Shannon entropy as a measure of lexical diversity in the texts:

$$H(text) = -\sum_{x \in text} \frac{freq(x)}{len(text)} log_2(\frac{freq(x)}{len(text)})$$

Here, x stands for all unique tokens/n-grams, freq stands for the number of occurrences in the text, and len for the total number of tokens in the text.

We added to the above features other features like punctuation, average sentence length, number of personal pronouns, and readability index to measure the complexity of a text. In punctuation, we included question marks, exclamation marks, double quotes, ellipses, and commas. To calculate a reading difficulty, we used textstat python package.

PoS tagging is another approach to finding informative features. We used the following PoS groups: nouns, verbs, infinitive verbs, adjectives, and different types of pronouns.

In our opinion, the number of proper nouns in the news should be higher because news articles describe more events and contain more details. For this reason, we look at the number of proper nouns in a text. We, therefore, look at numbers and Month-related words.We also introduced such features as a number of words-quantifiers like *everything*, *everyone*, *anyone*, *always*, *forever*, *never*, *constantly*, *nobody*, *nothing – vse*, *kazhdyj*, *ljuboj*, *vsegda*, *vechno*, *nikogda*, *postojanno*, *nikto*, *nichego*, and number of plural nouns.

Another set of measures is based on the idf – inverse document frequency for a word. Also, we experimented with the percentage of unique words used in different types of texts.

For many of our features, we used tokenization and lemmatization, and we used the morphological analyzer pymorphy2 for PoS tagging. We used basic stop words from NLTK library and add the most common words from our corpus. All features were normalized between 0 and 1 by the Min-Max scaling algorithm.

Finally, we defined the next groups of features: complexity features that include lexical diversity, lexical density, punctuation, average sentence length, number of personal pronouns, readability index, detail features that include the number of proper nouns in the text, numbers, and month-related words, imaginative features that include different PoS tags, words-quantifiers, and plural nouns.

Algorithms

We have chosen two machine learning algorithms: k++ means algorithm to categorize texts and the Random Forest algorithm to classify texts.

K-means++ algorithm

We use k++ means algorithm using the features described above to categorize texts. By using the feature weights, we will be able to obtain a ranking of the feature importance with regard to each class.

Let $X = \{x_i\},\ i = \overline{1, n}$ be the set of n d-dimensional points to be clustered into a set of K clusters, $C = \{c_k\}, k = \overline{1, K}$. K-means algorithm finds a partition of the set of dimensional points into a set of clusters such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let μ_k be the mean of cluster C_k. The squared error between μ_k and the points in cluster C_k is defined as

$$J(C_k) = \sum\nolimits_{x_i \in C_k} \|x_i - \mu_k\|^2$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(C) = \sum_{k=1}^{K} \sum\nolimits_{x_i \in C_k} \|x_i - \mu_k\|^2$$

In such a case, the general procedure is to search for a K-partition with the locally optimal within-cluster sum of squares by moving points from one cluster to another.

In our case, $X = \{x_i\},\ i = \overline{1, n}$ are parameters with raw values received using our approach. To normalize them, we use the formula:

$$z_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

where z_i is normalized data.

After normalizing the parameters, texts can be regarded as points in multidimensional space with the parameters as their coordinates. To divide the points, we (1) arbitrarily select an initial partition with K clusters; (2) generate a new partition by assigning each point to its closest cluster center; (3) compute new cluster centers and repeat steps 2 and 3 until cluster membership stabilizes.

Random Forest algorithm

Random forests or random decision forests technique is an ensemble learning method for text classification. To test this classifier, we use a standard 10-fold cross-validation experimental setup. This means that we randomly split the data into 10 equally sized chunks, and use 9 of those chunks to train the classifier.

By using the Random Forest classification, we will be able to obtain a measure of the importance of the predictor features. This is a difficult concept to define in general, because the importance of a variable may be due to its (possibly complex) interaction with other variables. The random forest algorithm estimates the importance of a feature by looking at how much prediction error increases when data for that feature is permuted while all others are left unchanged. The necessary calculations are carried out tree by tree as the random forest is constructed. Table 1 shows the result for the two algorithms.

Experiments and Evaluation

We use two algorithms for distinguishing texts: k-means++ for cluster analysis and Fandom Forest classifier as described above. We train with a Random Forest classifier with each set of features described above and evaluate the predictions using 10-fold cross-validation. We experiment with all features, with complex features, imaginative, and detailed features. The accuracy of predictions is shown in Table 1.

Table 1. Accuracies of differentiating texts

Features	*Accuracies*	
	Random Forest	*k-means++*
Complexity	*0.854*	*0.677*
Imaginative	*0.596*	*0.551*
Detailed	*0.426*	*0.553*
All features	*0.870*	*0.685*

Also, we have experimented with individual features to distinguish texts using k-means++ algorithm for all features. The best accuracy is obtained with text length (0.663), lexical diversity (0.619), question marks (0.605), and nouns (0.547). Ellipses are the worst feature with only 0.509 accuracy.

We take a closer look at the importance of the features. We have used the Random Forest algorithm to obtain a ranking of the features. Our experiments are shown the same results as the experiments with individual features. The most important feature is obtained with text length (0.199), lexical diversity (0.166), question marks (0.165), and nouns (0.043). Ellipses are a less important feature (0.013).

The clusterization results using k-means++ algorithm and the measure of feature importance by Random Forest can be seen in Table 2.

Table 2. Accuracies of k-means++ clusterization and Measure of Feature Importance of Random Forest

Features	*Accuracies*	*Measure of Feature Importance*
Nouns	*0.547*	*0.043*
Verbs	*0.519*	*0.037*
Adjectives	*0.531*	*0.040*
Plural nouns	*0.510*	*0.022*
Question marks	*0.605*	*0.165*
Exclamation marks	*0.513*	*0.018*
Quotes	*0.528*	*0.038*
Ellipses	*0.509*	*0.013*
Commas	*0.512*	*0.024*
Numbers	*0.518*	*0.035*
Quantifiers	*0.522*	*0.033*
Demonstrative pronounces	*0.516*	*0.028*
Pos pronounces	*0.519*	*0.021*
Proper nouns	*0.518*	*0.040*
Average sentence length	*0.510*	*0.041*
Lexical diversity	*0.619*	*0.166*
Lexical density	*0.521*	*0.041*
Text length	*0.663*	*0.199*

For some important features like diversity, text length, question marks, and nouns is interesting how they are distributed. To this, boxplots for these features are provided in Figure 1.

Then we experiment with various feature combinations to obtain combinations that worked best. We use k-means++ algorithm for this experiment. The combination of text length and number of question marks achieved an accuracy of 0.673; the combination of average sentence length, number of question marks, and the text length achieved an accuracy of 0.679; the combination of diversity, average sentence length, number of question marks, and the text length achieved an accuracy of 0.705; the combination of demonstrative pronounces, diversity, average sentence length, number of question marks, and the text length achieved an accuracy of 0.708: the combination of nouns, adjectives, questions, proper nouns, diversity, length text achieved an accuracy of 0.710; the combination of demonstrative pronounces, diversity, average sentence length, number of question marks, and the text length achieved an accuracy of 0.708; the combination of nouns, adjectives, questions, average sentence length, demonstrative pronounces, diversity, length text achieved an accuracy of 0.710; the combination of nouns, questions, exclamation marks, possessive pronouns, average sentence length, diversity, lexical density, and length text achieved an accuracy of 0.715; the combination of nouns, questions, quantifiers, demonstrative pronounces, average sentence length, diversity, lexical density, and length text achieved an accuracy of 0.716.

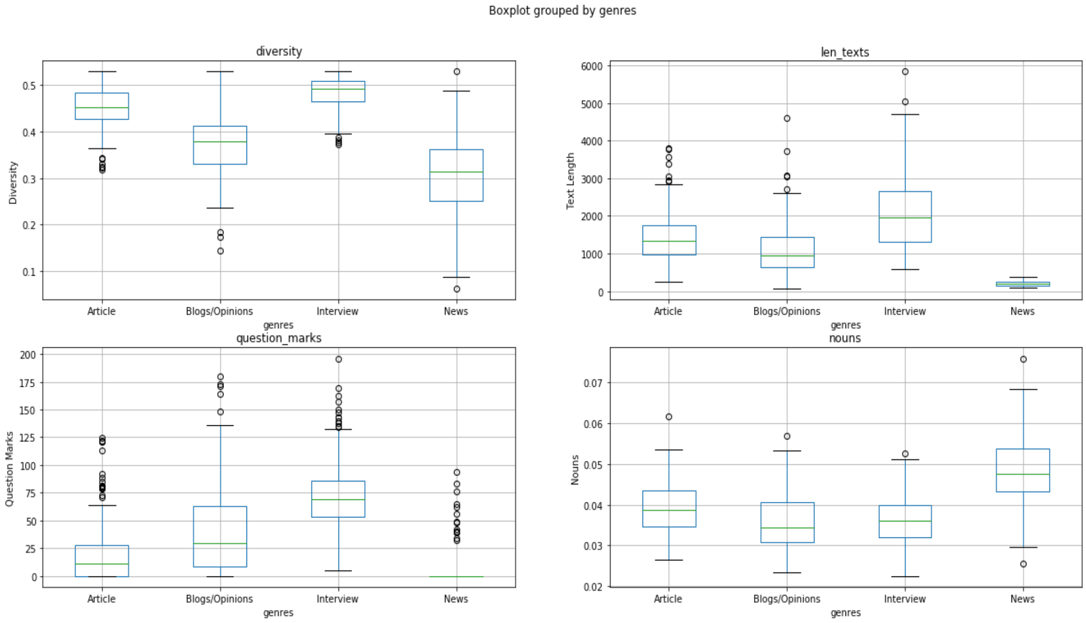

Fig. 1. Boxplots of more important features.

The best accuracy has been obtained with a combination of 10 features (nouns, adjectives, plural nouns, question marks, numbers, quantifiers, proper nouns, average sentence length, diversity, lexical density, and length text), 0.719. The accuracy is grown from a combination of two to ten features, while it is beginning to fall from a combination of 11 features to all features.

Figure 2 shows the genre distribution of some notable features. The analysis of the data we have received enabled the following conclusions to be drawn: the hypothesis about the importance of text length for genres has been verified; the hypothesis about the importance of complexity of interview genre has been verified; the hypothesis about the importance of proper nouns for genres has been not verified; the hypothesis about the importance of numbers for interview and blogs genres has been verified.

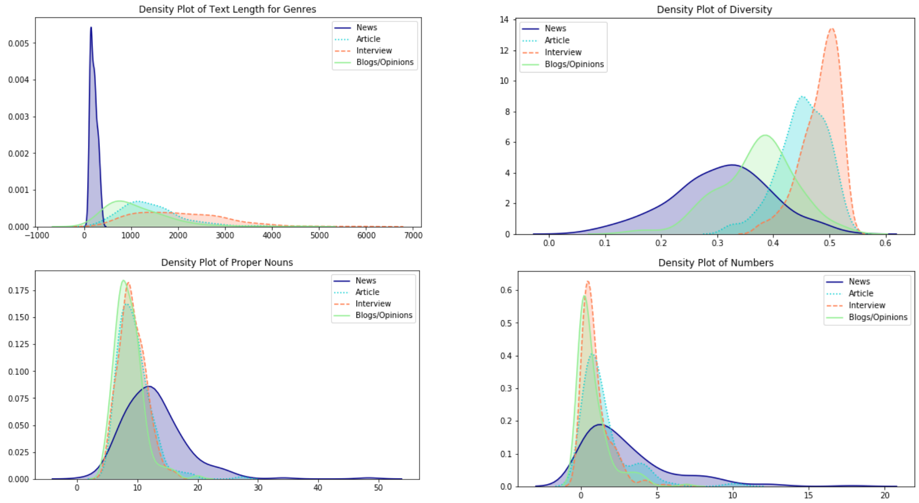

Fig. 2. Density plots of notable features.

Also, we have taken a closer look at the words that are associated with different types of texts. We have analyzed the words that are most commonly used in all types of texts and selected the ones that are unique for each type of text. We have identified words with the highest weight in the Random Forest model. The table shows the words that are unique to each type of text.

Table 2. Top unique selected words in different types of texts

Type	*Top selected unique words*
News	*media, agency, user, victim, attack, police, press, management, car, service, driver, version, blow, operation, data, object, accident, akim, leader, military, communication, freedom, provision, expert, court, death, attempt, violation, production, statement, portal, article, committee, composition, shop, publication, department, match, incident, mode, square, message, sale, chairman, deputy, ministry of internal affairs, incident* *smi, agentstvo, pol'zovatel', postradavshij, napadenie, policija, pressa, rukovodstvo, avtomobil', sluzhba, voditel', versija, udar, operacija, dannye, ob#ekt, dtp, akim, lider, voennyj, kommunikacija, svoboda, obespechenie, jekspert, sud, smert', popytka, narushenie, proizvodstvo, zajavlenie, portal, stat'ja, komitet, sostav, magazin, izdanie, vedomstvo, match, proisshestvie, rezhim, ploshhad', soobshhenie, prodazha, predsedatel', zamestitel', mvd, incident* https://translit.ru/
Blogs	*aul, status, position, patient, main, essence, character, spirit, disease, gift, door, coma, nature, ancestor, set, circle, list, custom, heart, thought, genus, kazakh, tradition, steppe, member, loved ones, god, head, generation, medicine, century, frame, happiness, picture, answer, light, earth* *aul, status, polozhenie, pacient, glavnoe, sut', harakter, duh, bolezn', podarok, dver', koma, priroda, predok, mnozhestvo, krug, spisok, obychaj, serdce, mysl', rod, kazah, tradicija, step', blizkie, bog, golova, pokolenie, medicina, vek, kadr, schast'e, kartina, otvet, svet, zemlja* https://translit.ru/
Opinions	*watch, salary, housing, class, China, Russia, Bank, capital, birth, phone, analysis, option* *chasy, zarplata, zhil'jo, klass, Kitaj, Russia, bank, stolica, rozhdenie, telefon, analiz, variant* https://translit.ru/
Interview	*need, understanding, creature, proposal, base, fund, practice, indicator, basis, astana, technology, stage, requirement, training, culture, environment, university, science, preparation, position, knowledge, approach, sport, factor, implementation, team, product* *neobhodimost', ponimanie, sozdanie, predlozhenie, baza, fond, praktika, pokazatel', osnova, astana, tehnologija, jetap, trebovanie, obuchenie, kul'tura, sreda, universitet, nauka, podgotovka, pozicija, znanie, podhod, sport, faktor, realizacija, komanda, product* https://translit.ru/

Words such as *policija*, *pressa*, and *akim* acted as identifiers of news articles. It is expected that such words as *aul*, *status*, *polozhenie*, *step'*, *pokolenie*, etc. became unique for blogs, whereas in articles such words as *Kitaj*, *Russia*, *bank*, *housing* and others most often meet. In the interview there are the words *neobhodimost'*, *ponimanie*, *sozdanie*, *kul'tura*, *sreda*, *universitet*, *nauka*.

Conclusion

The number of complex documents, including online news, is growing every year. Along with information-dense publications like news, the individual and personified texts like articles, blogs, and interviews are published. These types of texts might be opinionated. In order to distinguish different types of news or analyze them, various kinds of algorithms and methods are needed, including methods for extracting the most informative features.

We have presented a feature-based language-independent approach to distinguish the genres like news, articles, interview, blogs/opinions. We analyzed the publications collected from the official online news portal. Our corpus contained four genres: news, articles, interviews, blogs/opinions. We hypothesized that there are a set of features that distinguish the genres with a high degree of accuracy. We proposed three groups of features for detecting the genres. Our experiments have shown that our model can distinguish genres with high accuracy.

References:

1 B. Kessler, G. Numberg, and H. Schütze, “Automatic detection of text genre,” 1997.

2 F. Morstatter, L. Wu, U. Yavanoglu, S. R. Corman, and H. Liu, “Identifying Framing Bias in Online News,” ACM Trans. Soc. Comput., vol. 1, no. 2, pp. 1–18, 2018.

3 A. Louis and A. Nenkova, “A corpus of general and specific sentences from news,” Proc. 8th Int. Conf. Lang. Resour. Eval. Lr. 2012, pp. 1818–1821, 2012.

4 M. Hardalov, I. Koychev, and P. Nakov, “In search of credible news,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.

5 J. J. Li and A. Nenkova, “Fast and accurate prediction of sentence specificity,” Proc. Natl. Conf. Artif. Intell., vol. 3, pp. 2281–2287, 2015.

6 V. A. Yatsko, M. S. Starikov, and A. V. Butakov, “Automatic genre recognition and adaptive text summarization,” Autom. Doc. Math. Linguist., vol. 44, no. 3, pp. 111–120, 2010.

7 T. Portele, “Data-driven classification of linguistic styles in spoken dialogues,” 2002.

8 E. N. Forsyth and C. H. Martell, “Lexical and discourse analysis of online chat dialog,” in ICSC 2007 International Conference on Semantic Computing, 2007.

9 D. Gîfu and D. Cristea, “Monitoring and predicting journalistic profiles,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013.

10 V. Johansson, “Lexical diversity and lexical density in speech and writing: a developmental perspective,” Work. Pap. Linguist., 2009.

11 W. DuBay, “The Principles of Readability.,” Online Submiss., 2004.

12 S. Oraby, L. Reed, S. Tandon, S. T.S., S. Lukin, and M. Walker, “Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators,” 2019.

13 O. Dušek, J. Novikova, and V. Rieser, “Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge,” Comput. Speech Lang., 2020.

14 J. Novikova, A. Balagopalan, K. Shkaruta, and F. Rudzicz, “Lexical Features Are More Vulnerable, Syntactic Features Have More Predictive Power,” 2019.