УДК: 00.007
DOI: 10.25559/SITITO.16.202001.90-98

Особенности формирования корпуса текстов казахстанского сегмента электронных новостей

У. А. Оспанова, М. А. Баймаханбетов, И. Г. Акоева*, Т. К. Булдыбаев, М. К. Атанаева
АО «Информационно-аналитический центр» МОН РК, г. Нур-Султан, Казахстан
010000, Республика Казахстан, г. Нур-Султан, пр. Достык, д. 18
* inessa.akoyeva@iac.kz

Аннотация

Культура потребления населением онлайн-новостей продолжает формироваться и набирает популярность, увеличивая аудиторию читателей. В тоже время возрастает и число тех, кто попадает под негативное влияние недостоверных новостей. Перед исследователями встает задача анализа массмедиа. Одним из направлений анализа новостного контента является тематическое моделирование, распознавание фейковых новостей, сентимент-анализ. Однако для проведения исследований в данных областях необходим размеченный корпус.

В данной работе представлены методологические основы формирования корпуса описывающих сбор данных и отбор источников для формирования корпуса. Также представлено описание теоретических основ представительности и сбалансированности и обосновано соответствие корпуса выдвигаемым требованиям. В ходе проведенного комплекса работ сформирован корпус новостных текстов, включающий 1,9 млн статей из 22 источников. Проведена разметка корпуса. Проведен анализ тематической структуры сформированного корпуса с применением модели LDA.

Сформированный корпус позволит апробировать алгоритмы машинного обучения, направленные на распознавание отдельных информативных признаков, и выявить закономерности, присутствующие в массиве новостных публикаций. Также корпус будет полезен исследователям в области машинного обучения и NLP для тестирования алгоритмов машинного обучения согласно собственным целям.

Ключевые слова: корпус, разметка, тональность, объективность, СМИ, информативные признаки.

Финансирование: данная статья подготовлена в рамках реализации ПЦФ № BR05236839 по теме «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана» при поддержке Комитета науки Министерства образования и науки Республики Казахстан.

Для цитирования: Оспанова, У. А. Особенности формирования корпуса текстов казахстанского сегмента электронных новостей / У. А. Оспанова, М. А. Баймаханбетов, И. Г. Акоева, Т. К. Булдыбаев, М. К. Атанаева. – DOI 10.25559/SITITO.16.202001.90-98 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 1. – С. 90-98.

Features for Forming Text Corpus of Kazakhstan Electronic News

U. A. Ospanova, M. A. Baimakhanbetov, I. G. Akoyeva*, T. K. Buldybayev, M. K. Atanayeva
Information-Analytical Center, JSC, Nur-Sultan, Kazakhstan
18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan
* inessa.akoyeva@iac.kz

Abstract

The culture of online-news consumption continues to take shape and is gaining popularity, increasing the audience of readers. At the same time, the number of those who fall under the negative influence of false news is growing. Researchers are faced with the task of analyzing mass media. One of the areas of news content analysis is thematic modelling, recognition of fake news, sentiment analysis. However, to research these areas, there is a need in a labelled corpus.
This paper presents the methodological foundations of the corpus formation. It describes the process of data collection and the selection of sources to form the corpus. It also presents a description of the theoretical foundations of representativeness and balance and explains compliance of the corpus with the requirements. In the course of the composite work, authors gained a corpus of 1.9 million news texts from 22 news sources. They conducted corpus markup and carried-up the analysis of the thematic structure of the formed corps using the LDA model.
The formed corpus will allow testing machine learning algorithms aimed at recognizing individual informative features and identifying patterns that are present in the array of news publications. Also, the corpus will be useful to machine learning and NLP researchers to test machine learning algorithms according to their own goals.

Keywords: corpus, markup, sentiment, objectivity, Mass Media, informative features.

Funding: This article was prepared as part of the implementation of the FDF No. BR05236839 on the topic "Development of Information Technologies and Systems to Stimulate the Sustainable Personal Enhancement as One of the Foundations for the Development of Digital Kazakhstan" with the support of the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan.

For citation: Ospanova U.A., Baimakhanbetov M.A., Akoyeva I.G., Buldybayev T.K., Atanayeva M.K. Features for Forming Text Corpus of Kazakhstan Electronic News. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(1):90-98. DOI: https://doi.org/10.25559/SITITO.16.202001.90-98

Введение

Сложившаяся культура потребления новостей электронных СМИ [1], использования социальных сетей и мессенджеров обмена мгновенными сообщениями помимо очевидной пользы, порой невольно, а иногда целенаправленно способствует распространению недостоверной информации. Необходимость своевременного реагирования на распространение такого рода информации требует поиска способов минимизации влияния недостоверной и искаженной информации на социум. Выделение общих характеристик и паттернов в структуре формирования резонансных новостей в некоторой степени поможет в решении данного вопроса [2-4]. Для разработки и тестировании алгоритмов машинного обучения необходим корпус текстовой информации. В Казахстане исследования в области корпусной лингвистикой или разметки корпусов мало представлены[1]. Таким образом, вопрос формирования объемного корпуса новостных текстов для казахстанского сегмента является актуальным.

Одним из самых важных требований при формировании корпуса является обеспечение представительности (репрезентативности), под которой понимается способность корпуса отражать все свойства исследуемой области [5].

Д. Бибер подчеркивает, что при формировании корпусов необходимо получить максимальную вариативность данных, описывающих исследуемый феномен для более глубокого его понимания [6].

Корпусная лингвистика опирается на корпус как на выборку большей популяции. Как и во всех других научных дисциплинах, эмпирические результаты, полученные на основе выборки, могут быть обобщены для большей популяции только в том случае, если выборка является репрезентативной для этой популяции.

Сам опыт решения вопроса представительности описан в работах С.Т. Грайса [7,8], Г. Лича [9], Т. Варади [10], которые уделили внимание рекомендациям решения вопроса репрезентативности. Однако существует альтернативное мнение, согласно которому представительность корпуса не является существенной и обеспечивается за счет увеличения количества текстов в корпусе, т.е. объема выборки. Апологетами этой точки зрения являются Дж. Синклэр [11] и П. Хэнкс [12].

Также в работе Мак Энери и соавторов [13,14] описывается, что концептуально и методологически существует два основных типа репрезентативности в отношении двух параметров: исследуемая область и лингвистическая репрезентативность.

Репрезентативность корпуса в отношении исследуемой области представляет собой ту степень, в которой корпус содержит полный диапазон вариативности по типам тестов, характерным исследуемой области. Репрезентативность в отношении исследуемой области является решающим аспектом для исследователя, и позволяет ему проецировать результаты, полученные на основе анализа одной конкретной выборки на всю исследуемую область.

Лингвистическая репрезентативность представляет собой ту степень, в которой корпус содержит весь полный диапазон вариативности лингвистических характеристик, присущих исследуемой области. Лингвистическая репрезентативность определяет пригодность выборки корпуса для ответа на конкретные вопросы исследования о конкретных языковых особенностях. Важно отметить, что языковая репрезентативность по своей сути связана с изучаемым языковым уровнем; один и тот же корпус может быть представителем общей грамматической структуры, но не лексических распределений.

Основной методологической проблемой при оценке репрезентативности корпуса является оценка языковых и ситуационных характеристик цензуса, с которым сравнивают корпус. Поскольку собрать всю совокупность данных, отражающих определенный феномен, как правило, не представляется возможным.

Б. Грей и соавторы в своей работе, посвященной методам оценки представительности корпуса обобщили подход к такой оценке, принятый в научно-исследовательской деятельности[2] [15] в соответствии с представлением в Таблице 1, где отображены параметры оценки представительности (1 и 2), исходя из возможности (A) или невозможности (B) определения характеристик цензуса, а также сценарии оценки представительности корпуса в этих случаях.

Таблица 1. **Подходы к оценке репрезентативности корпуса**
Table 1. **Approaches to assessing the corpus representativeness**

	A. Возможность определения характеристик цензуса присутствует	**B. Возможность определения характеристик цензуса отсутствует**
1. Представительность в отношении исследуемой области	**A1.** Проведение детального анализа ситуационных характеристик каждого текста в корпусе. Сравнение таких характеристик с характеристиками исследуемой области. Другим вариантом является экстраполяция обобщений по корпусу только на определенную часть популяции.	**B1.** Сравнение встречаемости и / или пропорцию ситуационных характеристик, представленных в полной популяции, с серией экспериментальных под-корпусов, которые были собраны при помощи различных методов формирования корпусов.
2. Лингвистическая представительность	**A2.** Разделение сформированного корпуса на более мелкие подмножества, отобранные методом случайной выборки. Сравнение дисперсии лингвистических характеристик в малых выборках с лингвистическими характеристиками всего корпуса.	**B2.** Сравнение распределения ряда лингвистических особенностей цензуса, с серией экспериментальных под-корпусов, которые были собраны при помощи различных методов формирования корпусов.

[1] Национальные ресурсы НТИ / Отчеты о НИР // Национальный научный портал Республики Казахстан [Электронный ресурс]. URL: http://nauka.kz/page.php?page_id=108&lang=1 (дата обращения: 08.02.2020).

[2] Gray, B. Exploring Methods for Evaluating Corpus Representativeness / B. Gray, J. Egbert, D. Biber // Paper presented at the Corpus Linguistics International Conference 2017. – Birmingham, UK, 2017. – URL: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper277.pdf (дата обращения: 08.02.2020).

Под сбалансированностью корпуса понимается обеспечение пропорционального присутствия в корпусе текстов, отражающих категории исследуемой области. В сбалансированных корпусах в корпус включаются тексты, представляющие разные модусы дискурса, разнообразные по жанрам, стилям, тематике, классам и т.п.

В целом сбалансированность корпуса является качеством корпуса, достигаемым искусственным путем. Сбалансированность корпуса может достигаться за счет равноценного распределения каждой из категорий (одинаковое количество текстов, представляющих определенную категорию) или за счет пропорционального распределения[3]. В последнем случае каждой категории текстов присваивается определенная весовая доля, в соответствии с которой и определяется количество текстов в корпусе. Категория (признак, класс), по которой оценивается сбалансированность корпуса, зависит от целей исследования. Кроме того, в зависимости от них же, при обеспечении сбалансированности корпуса можно учитывать не одну, а несколько категорий.

Таким образом, сбалансированные корпуса имеют фиксированный объем, и процедура пополнения таких корпусов происходит согласно строгой и четко выработанной процедуре отбора. Также принцип сбалансированности может быть соблюден в разрезе источников (определенное пропорциональное соотношение публикаций из разных источников).

Составление тестового корпуса обусловлено необходимостью экспериментального подтверждения эффективности работы алгоритмов машинного обучения, методов и походов, выработанных для целей проекта, а также нецелесообразностью проведения тестирования на всем корпусе при машинной обработке данных и вычислительных процессах ввиду больших затрат времени.

При формировании корпуса были учтены существующие практики работы с текстовыми корпусами. Изученная информация была структурирована по основным характеристикам, необходимым для учета в дальнейшей работе.

Методология и данные

Для формирования корпуса с помощью парсеров произведен сбор 1,9 млн текстовых статей из 22 открытых информационных интернет-источников, затем произведены предварительная обработка корпуса и его метаразметка.

Оценка содержания публикаций по степени тональности и объективности произведена с использованием лингвистических словарей, сформированных в рамках исследования совместно с экспертами-лингвистами.

В нашем случае репрезентативность в отношении исследуемой области корпуса не подлежит сомнению, так как выборка корпуса включила практически весь цензус по источникам, отобранным методом случайной выборки (т.е. всю популяцию текстов открытых информационных источников, которые могут оказывать потенциальное влияние на социум). Таким образом, главное условие формирования корпуса, как отражение объективной реальности, полностью соблюдено.

Для обеспечения универсальности корпуса при многократном использовании существуют рекомендованные стандарты предоставления метаданных разметки текстовых корпусов, на основе Text Encoding Initiative (TEI) и EAGLES (Expert Advisory Group on Language Engineering Standards). В качестве языка разметки используются универсальные языки разметки данных, такие как, например, SGML и XML [16].

С целью соблюдения всех методологических условий, которые позволяют экстраполировать результаты, полученные на тестовом корпусе на всю совокупность данных в составе корпуса отбор публикаций в тестовый корпус производился методом случайной выборки от общего объема новостей из 22 источников, при котором каждая публикация имела одинаковые шансы попасть в состав формируемого корпуса.

Вместе с тем необходимо отметить, что отсутствие на некоторых интернет-ресурсах возможности отставить комментарий к публикации значительно усложняет анализ данных, поскольку с точки зрения оценки влияния на социум количество комментариев и просмотров публикаций является, почти единственным прямым индикатором заинтересованности населения [17-20].

Для целей исследования наличие этих показателей является критичным. По этой причине из списка источников, вошедших в корпус, были отобраны те, которые содержат минимальный набор данных, позволяющих произвести оценку интереса казахстанского общества к теме, поднимаемой в публикации: количество комментариев и количество просмотров. Итоговое количество текстов, вошедших в тестовый корпус, соответствующий всем необходимым параметрам, составило 35 643 публикации.-

1. Разметка корпуса

1.1 Признаки

Сформированный корпус был размечен (1) словарными признаками, (2) лексическими признаками, (3) признаками объема документа, (4) частями речи (PoS tagging), (5) тематическими областями, (6) информативными признаками содержания статей. Словари по тематическим областям и информативным признакам были составлены на основе углубленного лингвистического анализа 2000 новостных текстов, отобранных методом случайной выборки. Рассмотрим данные группы признаков.

1) Словарные признаки

Эта группа объединяет N-граммы, рассчитанные на основе встречаемости количества уникальных слов (1) и биграмм (2). Другие 4 предиктора в этой группе были получены во время процедуры text-mining с помощью программного обеспечения Statistica[4], которое представляет собой автоматическую обработку текстовой информации, приводящую к цифровому представлению документов путем определения нескольких наиболее важных измерений (называемых «концептами»), которые имеют наибольшую вариативность между словами и документами – 4 концепта (3-6).

2) Лексические признаки

Большинство лексико-семантических признаков также были получены с помощью подхода «мешок слов» и включают количество точек (7), запятых (8), восклицаний (9), вопросов (10), многоточий (11), кавычек (12), скобок (13), тире (14) и

[3] Corpus Linguistics. Method, Theory and Practice [Электронный ресурс]. URL: http://corpora.lancs.ac.uk/clmtp/1-data.php (дата обращения: 08.02.2020).

[4] StatSoft [Электронный ресурс]. URL: http://statsoft.ru (дата обращения: 08.02.2020).

дефисов (15). Два других предиктора, подпадающих под эту категорию, были получены с помощью словарного подхода и представляют собой встречаемость слов из словаря тональности в документах – позитивные (16) и негативные (17) слова.

3) Признаки объема документа

Предикторы этой группы включают символы (18), слова (19) и предложения (20). Количество символов было рассчитано с учетом пробелов.

2.4) Части речи (PoS tagging)

Эта группа объединяет количество прилагательных (21), предлогов (22)[5], наречий (23), вспомогательных слов (24), координирующих союзов (27), определителей (26), междометий (27), существительных (28), чисел, включая цифры (29), частиц (30), местоимений (31), имен собственных (32), знаков препинаний (33), подчинительных союзов (34), глаголов (35). Для проведения PoS tagging была использована библиотека UDPipe[6], свободно доступная для языка статистического программирования R.

2.5) Тематические области

Тематические области были вычислены на основе встречаемости слов из словарей тематических сфер. Они включали словари по таким темам, как административно-территориальные вопросы (123 слова), экономика и финансы (238 слов), внешняя политика (95 слов), гражданское общество (188 слов), здравоохранение (141 слово), индустриально-инфраструктурное развитие и технологии (115 слов), оборона и безопасность (77 слов), образование и наука (160 слов), правительство и парламент (97 слов), происшествия и преступления (326 слов), сельское хозяйство (61 слово), социальное обеспечение (28 слов), спорт, культура и развлечения (243 слова), труд (25 слов), энергетика (70 слов), экология и туризм (40 слов), чрезвычайные ситуации (53 слова). Перечень тематических областей был составлен на основе анализа тематических тегов, заданных новостными сайтами, с последующим проведением экспертных консультаций (7 экспертов в области СМИ). Эта группа предикторов объединяет 17 соответствующих признаков (36-49), рассчитанных по встречаемости слов из этих словарей.

2.6) Информативные признаки контента документа

Данные характеристики были измерены с помощью 4 предикторов: политизированные (50) и манипулятивные (51) слова, слова, призывающие к действию (52) и предвзятые слова (53). Данные признаки рассчитаны на основе встречаемости слов, включенных в соответствующие словари, в соответствии с вышеупомянутым углубленным лингвистическим анализом.

1.2 Тональность

На данных тестового корпуса был проведен комплекс работ по его разметке с помощью сформированных словарей. Алгоритм разметки не требует машинного обучения, однако зависит от качества словаря позитивных и негативных слов.

Полярность тональности. Перед проведением алгоритма разметки была применена процедура распознавания частей речи (PoS tagging) на тестовом корпусе. Процедура PoS tagging затруднена для русского языка малым количеством надежных библиотек и открытых программных решений обученных словарей. Из существующих вариантов, выбор был сделан в пользу библиотеки Udpipe, свободно доступной для статистического программного обеспечения C++ и R.

После процедуры PoS tagging тексты были размечены с помощью словаря тональности. Цель состояла в том, чтобы получить индикатор тональности в диапазоне [-1:1], который затем был использован в качестве непрерывной или категориальной переменной. В случае категориальной переменной можно либо использовать бинарную переменную, либо назначить несколько категорий, разделив диапазон на 11 групп, причем средняя группа является нейтральной категорией. Полярность документа – это нормализованная сумма полярностей на уровне предложений, тогда как полярность на уровне предложений была рассчитана с использованием отрицателей, усилителей и уменьшителей.

Отрицатели использовались, чтобы изменить полярность слов, если отрицатель был найден до и после 4 слов тональности, он автоматически изменит знак слова тональности с +1 до -1. Рассмотрим следующее простое предложение:

Погода была плохой (без частицы-отрицателя
тональность предложения (-1)

Погода была не плохой, а... (с частицей-отрицателем,
тональность предложения (+1)

Не считая отрицателя «нет», можно ошибочно считать второе предложение позитивным, но присутствие отрицателя явно меняет полярность тональности. Таким же образом были использованы усилители и уменьшители, чтобы максимально точно определить интенсивность тональности.

Погода была весьма неплохой (с увеличителем очень
итоговая тональность (+1.5)

Алгоритм очень точно определяет усилитель и присваивает вес 0,5, точно так же алгоритм идентифицирует слова, которые уменьшают полярность слов в каждом предложении. Таким образом, результирующая полярность, полученная на уровне предложения, суммировалась на уровне документа в диапазоне [-1: 1], где значение ближе к -1 интерпретируется как публикация с более негативной тональностью, значение ближе к 0 – как публикация с более нейтральной тональностью, а значение ближе к 1 означает публикацию с более положительной тональностью [21].

1.3 Объективность

Для определения объективности новостного текста предложен алгоритм определения степени объективности. Объективность понимается как объективно поданная информация, соответствующая критериям достоверности без субъективной точки зрения автора.

Для определения степени объективности статьи используются сформированные лингвистические словари: (1) словарь персоналий; (2) компетентные источники; (3) используемые связи при ссылке на источник.

[5] ADP: adposition. [Электронный ресурс]. URL: https://universaldependencies.org/u/pos/ADP.html (дата обращения: 08.02.2020).

[6] Wijffels, J. Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the "UDPipe" / J. Wijffels // "NLP" Toolkit [Электронный ресурс]. URL: https://cran.r-project.org/web/packages/udpipe/index.html (дата обращения: 08.02.2020).

Степень объективности рассчитывается по следующей формуле:

$$CO = \sum_{i=1}^{n} \frac{x_i}{y_i} \quad (1),$$

где
CO — степень объективности
x_i — количество объективных предложений
y_i — общее количество предложений

Алгоритм определения степени объективности строится на основе соответствий заранее определенным условиям и определяется на уровне предложений. Предложение считается объективным если оно отвечает, как минимум одному их следующих условий.

1. Наличие слова-связки и сайта (доверенные новостные сайты);
2. Наличие слова-связки и компетентного источника+ персоналии;
3. Наличие слова-связки и цифры (статистической информации).

Условия можно сложить чтобы получить не только бинарный показатель «объективное-субъективное предложение», но и степень объективности как категориальный показатель, состоящий из максимум 3 категорий, где вес каждого условия равен 0.33.

Предложение считается субъективным, если в нем не содержится ни одного из вышеупомянутых условий и включены местоимения и глаголы с личными окончаниями, указывающие на определенное лицо. Полученная объективность на уровне предложений может быть суммирована и представлена как отношение объективных к субъективным предложениям и нормализована на уровне публикации. Таким образом, степень объективности находится в интервале – [0;1], где значение ближе к 0 означает, что текст субъективен, ближе к 1 – объективен.

В ходе анализа корпуса авторами исследования было выявлено, что характеристики корпуса и публикаций достаточно сильно различаются в зависимости от источника, из которого была взята определенная публикация. Так, для некоторых сайтов количество просмотров выше 3000 является средним показателем, для других – рекордным. Та же тенденция наблюдается в отношении количества комментариев. В этой связи отражение фактических характеристик тестового корпуса представляется полезным для более глубокого понимания особенностей корпуса. Также в процессе распознавания тональности новостных текстов исследователи зачастую сталкиваются с проблемой несбалансированности разметки корпусов. Наблюдаемый перекос в разметке текстов может объясняться отражением объективной реальности, когда преобладание нейтральных публикаций обусловлено принципами журналистской этики. В свою очередь, преобладание публикаций с негативной тональностью над позитивной можно объяснить их большей популярностью у аудитории и способностью к вирусному распространению, как было доказано в исследовании Хансена и соавторов [22].

2. Тематическая структура корпуса

Для определения тематической представленности новостей в сформированном и размеченном корпусе применялись тематические модели. Большинство моделей разрабатываются на основе латентного размещения Дирихле LDA. Модель латентного размещения Дирихле (Latent Dirichlet Allocation) была предложена Дэвидом Блеем в 2003 году [23]. Это современный активно развивающийся вероятностный инструментарий, который применяется в задачах анализа данных[7] [24]. Для решения данной задачи тематические модели оказались полезными инструментами [25].

В контексте новостных средств массовой информации существуют два основных типичных варианта применения тематического моделирования: анализ повестки дня СМИ и анализ новостных фреймов. Само тематическое моделирование, как процесс, является одним из инструментов автоматизированного анализа корпуса новостных текстов и оценки их влияния на социум. Проведенная процедура продемонстрировала свою жизнеспособность по распределению текстов по тематикам.

Выводы

В данной работе представлены методологические основы формирования корпуса, описывающие сбор данных и отбор источников для формирования корпуса. Также представлено описание теоретических основ представительности и сбалансированности, обосновано соответствие корпуса выдвигаемым требованиям. В ходе проведенного комплекса работ сформирован корпус новостных текстов, включающий 35 643 статьи из 5 источников. Проведена разметка корпуса по тональности и объективности новостных текстов. Также проведен анализ тематической структуры сформированного корпуса с применением модели LDA. В корпусе новости представлены сбалансировано по различным тематическим направлениям. Наиболее объемными топиками в корпусе представлены новости, связанные с государством, происшествиями.

Анализ источников, вошедших в корпус, показал, что данные метаразметки неоднородны, и часть необходимых данных отсутствует, в связи с чем полные данные на этапе метаразметки по всем источникам удалось собрать по таким параметрам, как текст, URL-адрес и дата публикации. Это обусловлено тем, что архитектура сайтов, отобранных в выборку, не всегда предусматривает возможность сбора и наличие всех отобранных параметров мета-разметки (отсутствие данных о количестве просмотров либо возможности комментировать публикацию).

В ходе анализа корпуса, авторами исследования было выявлено, что характеристики корпуса и публикаций сильно разнятся в зависимости от источника, из которого была взята определенная публикация. В связи с чем представляет интерес вопрос работы с несбалансированными данными в корпусах.

Проведенный комплекс работ на основе состава, структуры, представительности, сбалансированности и полноты корпуса позволил провести анализ фактических характеристик корпуса для более глубокого понимания его особенностей.

Результаты работы по формированию проклассифицирован-

[7] Воронцов, К. В. Вероятностное тематическое моделирование [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (дата обращения: 08.02.2020).

ного и размеченного корпуса новостных текстов могут быть применены в области социальных наук, журналистики и информационных технологий.

Список использованных источников

[1] Newman, N. Digital News Report 2017 / N. Newman, R. Fletcher, A. Kalogeropoulos, D. A. L. Levy, R. K. Nielsen. – Reuters Institute for the Study of Journalism, Oxford, UK, 2017. – URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf (дата обращения: 17.12.2019).

[2] Jang, S. M. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis / S. M. Jang, T. Geng, J.-Y.Q. Li, R. Xia, C.-T. Huang, H. Kim, J. Tang J. – DOI 10.1016/j.chb.2018.02.032 // Computers in Human Behavior. – 2018. – Vol. 84. – Pp. 103-113. – URL: https://www.sciencedirect.com/science/article/pii/S0747563218300906?via%3Dihub (дата обращения: 17.12.2019).

[3] Colliander, J. "This is fake news": Investigating the role of conformity to other users" views when commenting on and spreading disinformation in social media / J. Colliander. – DOI 10.1016/j.chb.2019.03.032 // Computers in Human Behavior. – 2019. – Vol. 97. – Pp. 202-215. – URL: https://www.sciencedirect.com/science/article/pii/S074756321930130X?via%3Dihub (дата обращения: 17.12.2019).

[4] Vasoughi, S. The spread of true and false news online / S. Vasoughi, D. Roy, S. Aral. – DOI 10.1126/science.aap9559 // Science. – 2018. – Vol. 359, issue 6380. – Pp. 1146-1151. – URL: https://science.sciencemag.org/content/359/6380/1146 (дата обращения: 17.12.2019).

[5] Баранов, А. Н. Корпусная лингвистика / А. Н. Баранов // Введение в прикладную лингвистику. – М.: Едиториал УРСС, 2003. – С. 118.

[6] Biber, D. Representativeness in Corpus Design / D. Biber. – DOI 10.1007/978-0-585-35958-8_20 // Current Issues in Computational Linguistics: In Honour of Don Walker / ed. by A. Zampolli, N. Calzolari, M. Palmer. Linguistica Computazionale, vol. 9. – Springer, Dordrecht, 1994. – Pp. 377-407. – URL: https://link.springer.com/chapter/10.1007/978-0-585-35958-8_20 (дата обращения: 17.12.2019).

[7] Gries, S. Th. Exploring variability within and between corpora: some methodological considerations / S.Th. Gries. – DOI 10.3366/cor.2006.1.2.109 // – Corpora. – 2006. – Vol. 1, issue 2. – Pp. 109-151. – URL: https://www.euppublishing.com/doi/10.3366/cor.2006.1.2.109 (дата обращения: 17.12.2019).

[8] Gries, S. Th. Dispersions and adjusted frequencies in corpora / S. Th. Gries. – DOI 10.1075/ijcl.13.4.02gri // International Journal of Corpus Linguistics. – 2008. – Vol. 13, issue 4. – Pp. 403-437. – URL: https://www.jbe-platform.com/content/journals/10.1075/ijcl.13.4.02gri (дата обращения: 17.12.2019).

[9] Leech, G. New Resources, or Just Better Old Ones? The Holy Grail of Representativeness / G. Leech // Corpus Linguistics and the Web / ed. by M. Hundt, N. Nesselhauf, C. Biewer. – Rodopi, Amsterdam, 2007. – Pp. 133-149.

[10] Váradi, T. The Linguistic Relevance of Corpus Linguistics / T. Váradi // Proceedings of the Corpus Linguistics 2001 Conference / ed. by P. Rayson, A. Wilson, T. McEnery, A. Hardie, Sh. Khoja. – UCREL Technical Papers. – Lancaster University, UK, 2001. – No. 13. – Pp. 587-593. – URL: http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/varadi.pdf (дата обращения: 17.12.2019).

[11] Sinclair, J. Corpus, Concordance, Collocation / J. Sinclair. – Oxford, UK: Oxford University Press, 1991.

[12] Hanks, P. The Corpus Revolution in Lexicography / P. Hanks. – DOI 10.1093/ijl/ecs026 // International Journal of Lexicography. – 2012. – Vol. 25, issue 4. – Pp. 398-436. – URL: https://academic.oup.com/ijl/article-abstract/25/4/398/923874?redirectedFrom=PDF (дата обращения: 17.12.2019).

[13] McEnery, A. Corpus-Based Language Studies: An Advanced Resource Book / A. McEnery, R. Xiao, Y. Tono. – New York: Routledge, 2006.

[14] Xiao, Z. Situation Aspect as a Universal Aspect: Implications for Artificial Languages / Z. Xiao, A. McEnery. – DOI 10.22425/jul.2002.3.2.139 // Journal of Universal Language. – 2002. – Vol. 3, No. 2. – Pp. 139-177. – URL: https://www.sejongjul.org/archive/view_article?pid=jul-3-2-139 (дата обращения: 17.12.2019).

[15] Egbert, J. Designing and evaluating language corpora / J. Egbert, B. Gray, D. Biber. – Cambridge: Cambridge University Press, 2017.

[16] Lüdeling, A. Corpus Linguistics / A. Ludeling, M. Kytö. – Vol. 1. – Walter de Gruyter, Berlin, 2008.

[17] Oeldorf-Hirsch, A. Posting, commenting, and tagging: Effects of sharing news stories on Facebook / A. Oeldorf-Hirsch, S. S. Sundar. – DOI 10.1016/j.chb.2014.11.024 // Computers in Human Behavior. – 2015. – Vol. 44. – Pp. 240-249. – URL: https://www.sciencedirect.com/science/article/pii/S0747563214006232?via%3Dihub (дата обращения: 17.12.2019).

[18] Liu, Q. Understanding News 2.0: A framework for explaining the number of comments from readers on online news / Q. Liu, M. Zhou, X. Zhao. – DOI 10.1016/j.im.2015.01.002 // Information & Management. – 2015. – Vol. 52, issue 7. – Pp. 764-776. – URL: https://www.sciencedirect.com/science/article/pii/S0378720615000038?via%3Dihub (дата обращения: 17.12.2019).

[19] Tsagkias, M. News Comments: Exploring, Modeling, and Online Prediction / M. Tsagkias, W. Weerkamp, M. de Rijke. – DOI 10.1007/978-3-642-12275-0_19 // Advances in Information Retrieval. ECIR 2010 / ed. by C. Gurrin [et al.]. Lecture Notes in Computer Science. – Vol. 5993. – Springer, Berlin, Heidelberg, 2010. – Pp. 191-203. – URL: https://link.springer.com/chapter/10.1007%2F978-3-642-12275-0_19 (дата обращения: 17.12.2019).

[20] Chung, D. S. Interactive Features of Online Newspapers: Identifying Patterns and Predicting Use of Engaged Readers / D. S. Chung. – DOI 10.1111/j.1083-6101.2008.00414.x // Journal of Computer-Mediated Communication. – 2008. – Vol. 13, issue 3. – Pp. 658-679. – URL: https://academic.oup.com/jcmc/article/13/3/658/4582979 (дата обращения: 17.12.2019).

[21] Атанаева, М. К. Определение тональности и объектив-

ности новостных текстов словарным подходом / М. К. Атанаева, Т. К. Булдыбаев, У. А. Оспанова, И. Г. Акоева, К. С. Нурумов, М. А. Баймаханбетов // Научный аспект. – 2019. – Т. 3, № 3. – С. 296-308. – URL: https://elibrary.ru/item.asp?id=41388548 (дата обращения: 17.12.2019).

[22] Hansen, L. K. Good Friends, Bad News - Affect and Virality in Twitter / L. K. Hansen, A. Arvidsson, F. A. Nielsen, E. Colleoni, M. Etter. – DOI 10.1007/978-3-642-22309-9_5 // Future Information Technology / ed. by J. J. Park, L. T. Yang, C. Lee. – Communications in Computer and Information Science, vol. 185. – Springer, Berlin, Heidelberg, 2011. – Pp. 34-43. – URL: https://link.springer.com/chapter/10.1007/978-3-642-22309-9_5 (дата обращения: 17.12.2019).

[23] Blei, D. M. Latent Dirichlet Allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of Machine Learning Research. – 2003. – Vol. 3. – Pp. 993-1022. – URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (дата обращения: 17.12.2019).

[24] Vorontsov, K. V. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization / K. V. Vorontsov, A. A. Potapenko. – DOI 10.1007/978-3-319-12580-0_3 // AIST'2014, Analysis of Images, Social Networks and Texts. – Communications in Computer and Information Science (CCIS). – Vol. 436. – Springer International Publishing Switzerland, 2014. – Pp. 29-46. – URL: https://link.springer.com/chapter/10.1007/978-3-319-12580-0_3 (дата обращения: 17.12.2019).

[25] Korenčić, D. Document-based Topic Coherence Measures for News Media Text / D. Korenčić, S. Ristov, J. E. Najder. - DOI 10.1016/j.eswa.2018.07.063 // Expert Systems with Applications. – 2018. – Vol. 114. – Pp. 357-373. – URL: https://www.sciencedirect.com/science/article/pii/S0957417418304883 (дата обращения: 17.12.2019).

Поступила 17.12.2019; принята к публикации 10.03.2020; опубликована онлайн 25.05.2020.

Об авторах:

Оспанова Улжан Абаевна, проектный менеджер Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Республика Казахстан, г. Нур-Султан, пр. Достык 18), магистр менеджмента, ORCID: http://orcid.org/0000-0003-0427-7931, ulzhan.ospanova@iac.kz

Баймаганбетов Мухит Абилкасымович, главный аналитик Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Республика Казахстан, г. Нур-Султан, пр. Достык 18), ORCID: http://orcid.org/0000-0003-4978-2704, m.baimakhnbetov@gmail.com

Акоева Инесса Георгиевна, главный аналитик Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Республика Казахстан, г. Нур-Султан, пр. Достык 18), ORCID: http://orcid.org/0000-0002-1984-150X, aveoka@mail.ru

Булдыбаев Тимур Керимбекович, директор Департамента прикладных исследований и разработок, АО «Информационно-аналитический центр» МОН РК (010000, Республика Казахстан, г. Нур-Султан, пр. Достык 18), ORCID: http://orcid.org/0000-0001-8199-653X, timur.buldybayev@iac.kz

Атанаева Мираим Кажмухамбетовна, и.о. президента, АО «Информационно-аналитический центр» МОН РК, (010000, Республика Казахстан, г. Нур-Султан, пр. Достык 18), магистр государственного и местного управления, ORCID: http://orcid.org/0000-0002-3067-356X, miraim@mail.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Newman N., Fletcher R., Kalogeropoulos A., Levy D. A. L., Nielsen R.K. Digital News Report 2017. Reuters Institute for the Study of Journalism, Oxford, UK; 2017. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf (accessed 17.12.2019). (In Eng.)

[2] Jang S.M., Jang S.M., Geng T., Li J.-Y.Q., Xia R., Huang C.-T., Kim H., Tang J. A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis. *Computers in Human Behavior*. 2018; 84:103-113. (In Eng.) DOI: https://doi.org/10.1016/j.chb.2018.02.032

[3] Colliander J. "This is fake news": Investigating the role of conformity to other users" views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*. 2019; 97:202-215. (In Eng.) DOI: https://doi.org/10.1016/j.chb.2019.03.032

[4] Vasoughi S., Roy D., Aral S. The spread of true and false news online. *Science*. 2018; 359(6380):1146-1151. (In Eng.) DOI: https://doi.org/10.1126/science.aap9559

[5] Baranov A.N. *Vvedenie v prikladnuyu lingvistiku* [Introduction to Applied Linguistics]. Editorial URSS Publ, Moscow; 2003. p. 118. (In Russ.)

[6] Biber D. Representativeness in Corpus Design. In: A. Zampolli, N. Calzolari, M. Palmer (ed.) Current Issues in Computational Linguistics: In Honour of Don Walker. Linguistica Computazionale, vol. 9. Springer, Dordrecht; 1994. p. 377-407. (In Eng.) DOI: https://doi.org/10.1007/978-0-585-35958-8_20

[7] Gries S.Th. Exploring variability within and between corpora: some methodological considerations. *Corpora*. 2006; 1(2):109-151. (In Eng.) DOI: https://doi.org/10.3366/cor.2006.1.2.109

[8] Gries S.Th. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*. 2008; 13(4):403-437. (In Eng.) DOI: https://doi.org/10.1075/ijcl.13.4.02gri

[9] Leech G. New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. In: M. Hundt, N. Nesselhauf, C. Biewer (ed.) Corpus Linguistics and the Web. Rodopi, Amsterdam; 2007. p. 133-149. (In Eng.)

[10] Váradi T. The Linguistic Relevance of Corpus Linguistics. In: P. Rayson, A. Wilson, T. McEnery, A. Hardie, Sh. Khoja (ed.) Proceedings of the Corpus Linguistics 2001 Conference. UCREL Technical Papers, Lancaster University, UK; 2001. no. 13. pp. 587-593. Available at: http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/varadi.pdf (accessed 17.12.2019). (In Eng.)

[11] Sinclair J. Corpus, Concordance, Collocation. Oxford, UK: Oxford University Press; 1991. (In Eng.)
[12] Hanks P. The Corpus Revolution in Lexicography. *International Journal of Lexicography*. 2012; 25(4):398-436. (In Eng.) DOI: https://doi.org/10.1093/ijl/ecs026
[13] McEnery T., Xiao R., Tono Y. Corpus-Based Language Studies: An Advanced Resource Book. New York: Routledge; 2006. (In Eng.)
[14] Xiao Z., McEnery A. Situation Aspect as a Universal Aspect: Implications for Artificial Languages. *Journal of Universal Language*. 2002; 3(2):139-177. Available at: https://www.sejongjul.org/archive/view_article?pid=jul-3-2-139 (accessed 17.12.2019). (In Eng.)
[15] Egbert, J. Designing and evaluating language corpora / J. Egbert, B. Gray, D. Biber. – Cambridge: Cambridge University Press, 2017.
[16] Lüdeling A., Kytö M. Corpus Linguistics. vol. 1. Walter de Gruyter, Berlin; 2008. (In Eng.)
[17] Oeldorf-Hirsch A., Sundar S.S. Posting, commenting, and tagging: Effects of sharing news stories on Facebook. *Computers in Human Behavior*. 2015; 44:240-249. (In Eng.) DOI: https://doi.org/10.1016/j.chb.2014.11.024
[18] Liu Q., Zhou M., Zhao X. Understanding News 2.0: A framework for explaining the number of comments from readers on online news. *Information & Management*. 2015; 52(7):764-776. (In Eng.) DOI: https://doi.org/10.1016/j.im.2015.01.002
[19] Tsagkias M., Weerkamp W., de Rijke M. News Comments: Exploring, Modeling, and Online Prediction. In: C. Gurrin et al. (ed.) Advances in Information Retrieval. ECIR 2010. Lecture Notes in Computer Science, vol. 5993. Springer, Berlin, Heidelberg; 2010. p. 191-203. (In Eng.) DOI: https://doi.org/10.1007/978-3-642-12275-0_19
[20] Chung D.S. Interactive Features of Online Newspapers: Identifying Patterns and Predicting Use of Engaged Readers. *Journal of Computer-Mediated Communication*. 2008; 13(3):658-679. (In Eng.) DOI: https://doi.org/10.1111/j.1083-6101.2008.00414.x
[21] Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G., Nurumov K.S., Baimahanbetov M.A. Determination of the sentiment and objectivity of news texts vocabulary approach. *Nauchnyi aspekt*. 2019; 3(3):296-308. Available at: https://elibrary.ru/item.asp?id=41388548 (accessed 17.12.2019). (In Russ.)
[22] Hansen L.K., Arvidsson A., Nielsen F.A., Colleoni E., Etter M. Good Friends, Bad News - Affect and Virality in Twitter. In: J.J. Park, L.T. Yang, C. Lee (ed.) Future Information Technology. Communications in Computer and Information Science, vol. 185. Springer, Berlin, Heidelberg; 2011. p. 34-43. (In Eng.) DOI: https://doi.org/10.1007/978-3-642-22309-9_5
[23] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003; 3:993-1022. Available at: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed 17.12.2019). (In Eng.)
[24] Vorontsov K., Potapenko A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. In: D. Ignatov, M. Khachay, A. Panchenko, N. Konstantinova, R. Yavorsky (ed.) Analysis of Images, Social Networks and Texts. AIST 2014. Communications in Computer and Information Science, vol. 436. Springer, Cham; 2014. p. 29-46. (In Eng.) DOI: https://doi.org/10.1007/978-3-319-12580-0_3
[25] Korenčić D., Ristov S., Najder J.E. Document-based Topic Coherence Measures for News Media Text. *Expert Systems with Applications*. 2018; 114:357-373. (In Eng.) DOI: https://doi.org/10.1016/j.eswa.2018.07.063

Submitted 17.12.2019; revised 10.03.2020; published online 25.05.2020.

About the authors:

Ulzhan A. Ospanova, Project Manager of the Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan), Master of Management, ORCID: http://orcid.org/0000-0003-0427-7931, ulzhan.ospanova@iac.kz
Mukhit M. Baimakhanbetov, Chief Analyst of the Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan), ORCID: http://orcid.org/0000-0003-4978-2704, m.baimakhnbetov@gmail.com
Inessa G. Akoyeva, Chief Analyst of the Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan), ORCID: http://orcid.org/0000-0002-1984-150X, aveoka@mail.ru
Timur K. Buldybayev, Director of the Department of Applied Research and Development, Information-Analytical Center, JSC (18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan), ORCID: http://orcid.org/0000-0001-8199-653X, timur.buldybayev@iac.kz
Miraim K. Atanayeva, Acting President of the Information-Analytical Center, JSC (18 Dostyk Ave., Nur-Sultan 010000, Kazakhstan), Master of Public Administration, ORCID: http://orcid.org/0000-0002-3067-356X, miraim@mail.ru

All authors have read and approved the final manuscript.