УДК 004.048:004.021

П.И. Банокин, Е.Е. Лунёва, А.А. Ефремов

Оценка профилей пользователей-экспертов социальной сети

Рассмотрены характеристики поведения интернет-троллей и дано обоснование необходимости оценки поведения влиятельных пользователей социальной сети. Представлен процесс создания профилей поведения пользователей-экспертов социальной сети для оценки их компетентности и степени принадлежности к классу пользователей-троллей. Создание профиля производится путем анализа сообщений пользователя и преобразования в векторы числовых значений. Дано описание архитектуры ансамблевого классификатора, используемого для оценки эмоционального тона сообщений. Проведена кластеризация профилей пользователей.

Ключевые слова: ансамблевый классификатор, пользователь-эксперт, профиль поведения, интернет-тролль, социальная сеть.

doi: 10.21293/1818-0442-2019-22-2-61-66

В настоящее время социальные сети широко используются для высказывания мнений относительно фактов, людей, событий, продуктов или услуг. Анализ данных из социальных сетей позволяет решать широкий класс задач, в том числе при выполнении маркетинговых и социологических исследований, формировании политических прогнозов, а также применяется при оценке репутационных рисков человека или компании [1]. При этом сообщения некоторых пользователей в социальной сети могут быть направлены не столько на высказывание своего мнения, сколько на создание социальной провокации в публикациях, заинтересованных в своей большей узнаваемости и публичности. Наибольшей интерес среди таких пользователей вызывают так называемые интернет-тролли.

Интернет-тролли – это пользователи социальных сетей, действия которых направлены на искажение общественного мнения, включая возбуждение агрессии или провоцирование панического поведения [2]. Мотивами совершения деструктивных действий таких пользователей могут быть психические расстройства, личная неприязнь или выполнение заказов для создания репутационного и экономического ущерба. Для достижения своих целей тролли создают сообщения с ложными фактами и явными или скрытыми оскорблениями, используют технологии манипулирования мнением [3].

Деятельность интернет-троллей может затруднить получение полезной информации из социальных сетей.

Большая интенсивность потока данных из социальных сетей обусловливает необходимость автоматизированного анализа, который включает в себя анализ количественных показателей активности пользователей в сети, таких как количество репостов, комментариев, упоминаний и т.д., а также анализ текстов сообщений. При этом важной задачей является поиск пользователей-экспертов или лидеров в заданной предметной области [1]. Значимость задачи идентификации пользователей-экспертов обусловливается тем, что именно данные пользователи задают тренд в оценивании или отношении к определенному факту, событию и т.п. Среди найденных экспертов целесообразно определить пользователей, имеющих признаки поведения интернет-тролля. Также важной характеристикой найденного эксперта является предвзятость или непредвзятость мнения. Пользователь социальной сети со стойкими предубеждениями сохраняет свое явно выраженное отрицательное или положительное отношение к изучаемым явлениям в течение длительного времени. Эксперт с предвзятым мнением обращает меньше внимания на изменчивую природу изучаемого объекта и новые факты о нем. Классификация найденных экспертов по их поведению может повысить качество проводимых в дальнейшем исследований, в том числе социологических и маркетинговых.

Поведение пользователей можно оценить в ходе анализа текстов сообщений при помощи явных и неявных признаков. Явными признаками являются использование нецензурной лексики и оскорбительных выражений, а также устойчивое значение эмоционального тона. Неявные признаки могут включать в себя использование метафор, сравнений и маскировку нецензурных слов.

В работе [4] утверждается о возможности использования классификаторов текстовых данных для определения отдельных признаков поведения троллей. Для определения троллей могут быть использованы заранее установленные критерии поведения, среди которых можно выделить использование прописных букв, скорость ответа на комментарии, использование нецензурных слов [5]. Также возможным признаком поведения интернет-тролля является узкая специализация: его сообщения преимущественно посвящены одной теме, используется одинаковый набор хэш-тегов, эмоциональный тон сообщений ярко выражен и устойчив [5].

Целью работы является разработка подхода к созданию профилей поведения пользователей социальной сети Twitter и их последующего использования для выявления пользователей-троллей.

Классификация тестовых данных

Для классификации текстовых данных широкое применение нашли алгоритмы глубокого обучения на основе нейронных сетей. Часто используемыми архитектурами нейронной сети являются сверточная

сеть на основе многослойного персептрона и LSTM-сеть [6], которые в том числе используются для анализа эмоционального тона текста [7]. В зависимости от тренировочного набора данных, используемого способа кодирования текста точность алгоритмов варьируется от 76 до 95% [8].

Дополнительным способом повышения точности классификации является использование ансамбля классификаторов [9]. В состав ансамбля могут входить классификаторы, обученные на разных наборах данных или отличающиеся используемым алгоритмом классификации [9]. Ансамблевый классификатор формирует итоговую оценку путем пропорционального или непропорционального сложения оценок классификаторов [9]. Преимуществами ансамблевого классификатора являются возможность сочетать несколько алгоритмов машинного обучения, обновлять состояние ансамбля за счет добавления новых классификаторов по мере обновления обучающих выборок данных и исключения устаревших или неэффективных классификаторов.

Альтернативным способом повышения точности является совершенствование процесса предварительной обработки данных. В источнике [10] предлагается использовать расширенные векторные представления слов, дополненные значения символов, из которых состоит слово. Также для повышения точности используются механизмы внимания совместно с нейронной сетью LSTM [11].

Ансамблевая архитектура выбрана авторами в качестве базовой архитектуры классификации ввиду ее расширяемости и возможности адаптации в условиях быстрого изменения лексики текстов, созданных пользователями социальных сетей.

Процесс классификации пользователей

Предложенный процесс классификации пользователей состоит из следующих стадий:

1. Загрузка данных по заданной предметной области из социальной сети, которая выполняется на основе заданных ключевых слов или хэш-тегов.

2. Поиск пользователей экспертов в соответствии с подходом [12, 13], основанным на анализе социального графа с использованием метода Боргатти и алгоритма Кендалла–Уэя.

3. Создание профилей поведения для пользователей-экспертов.

4. Вычисление меры разнообразия для каждого пользователя на основе созданного профиля.

Профиль поведения пользователя-эксперта

Поведенческий профиль пользователя социальной сети является структурой данных, хранящей вычисленные характеристики поведения на основе коллекции выбранных сообщений пользователя и его дискуссий. Авторы предлагают способ создания профиля поведения пользователя путем обработки выборки его сообщений $\mathbf{M}=\{m_1,m_2,...,m_s\}$, где s – количество элементов в выборке сообщений пользователя. Пусть профиль пользователя включает вектор характеристик поведения $\mathbf{D}=(d_1,d_2,...,d_n)$, где n – количество характеристик. Каждая из характеристик является мерой разнообразия поведения пользователя. Предполагается, что эксперты с высоким значением меры разнообразия имеют более широкую сферу интересов и менее склонны к предвзятости. Профиль пользователя включает следующие характеристики:

1. d_1 – значение энтропии эмоционального тона сообщений пользователя. Энтропия вычисляется по формуле Шеннона $H=-\sum_i p_i \log_2 p_i$ [14] на основе коллекции значений эмоционального тона сообщений пользователя, полученных с использованием ансамблевого классификатора.

2. d_2 – индекс лексического разнообразия MLTD [15], вычисленный для выборки **M** сообщений пользователя. Если у пользователя отсутствуют собственные сообщения, индексу MLTD присваивается значение «0».

Также профиль сопровождается набором явных критериев поведения интернет-тролля, который можно представить в качестве вектора $\mathbf{t}=(t^{\langle 1\rangle}, t^{\langle 2\rangle}, ..., t^{\langle k\rangle})^T$, где k – количество критериев. Предлагается использовать следующие критерии:

1. t_1 – использование нецензурной лексики, значения «0» или «1».

2. t_2 – написание сообщений прописными буквами без учета аббревиатур. Значения «0» или «1».

Определение значения эмоционального тона

Для вычисления значения эмоционального тона предлагается использовать ансамблевый классификатор, построенный на основе двух классификаторов архитектуры CNN с символьным кодированием входных данных, и один классификатор архитектуры LSTM с кодированием входных данных в виде векторов слов GloVe [11]. Компоненты векторов слов, полученные в [16] в результате исполнения алгоритма GloVe на больших выборках текстовых данных, отражают эмоциональный тон, а также другие семантические характеристики слова. В табл. 1 приведена предварительная оценка эффективности классификаторов, проведенная авторами на тестовых выборках сообщений. Под эффективностью классификации понимается доля верно классифицированных сообщений.

Необходимо отметить, что для ансамбля классификаторов (см. табл. 1) не требуется тренировочный набор данных, так как он формируется из ранее обученных моделей классификаторов. При работе ансамблевого классификатора осуществляется проверка совпадений слов сообщения с коллекцией векторных представлений Glove. При недостаточном числе совпадений (менее 50%) используется только классификатор № 2.

За счет использования ансамблевой архитектуры точность классификации тестового набора из 100 сообщений превысила 77%, полученных авторами в работе [7] при использовании только LSTM-классификатора.

Таблица 1

Классификаторы текстовых данных

Классификатор	Архитектура нейронной сети	Тренировочный набор данных	Доля верно классифицированных сообщений, %
1	LSTM	Kaggle Sentiment140	72
2	CNN	Kaggle Sentiment140	78
3	CNN	Imdb Sentiment Dataset	76
4	Ансамбль классификаторов 1, 2, 3		87

Экспериментальный анализ

Для оценки эффективности подхода к созданию профилей поведения пользователей социальной сети Twitter была проведена серия из 20 экспериментов. Каждый эксперимент проводился на 180–320 заранее размеченных сообщениях из социальной сети Twitter. При этом каждая выборка состояла из сообщений пользователей-троллей и пользователей-экспертов с высоким уровнем компетентности. Разметка данных из социальной сети Twitter производилась авторами работы. Средняя эффективность классификатора составила 83,5%.

В табл. 2 приведены примеры сообщений, классифицированных верно и с ошибкой.

Для обозначения результатов классификации используется знак «+» для положительно классифицированных сообщений и «–» для отрицательно классифицированных сообщений. Ошибку классификации сообщения № 5 можно объяснить употреблением слов в переносном значении и отсутствием достаточного количества примеров подобных речевых оборотов в тренировочном наборе данных. Несмотря на ошибки отдельных классификаторов в примерах № 2–4, использование ансамбля классификаторов обеспечивает верный итоговый результат.

В качестве гипотезы H_0 выбрано утверждение о принадлежности пользователя к классу троллей, в качестве альтернативной гипотезы H_1 – о принадлежности пользователя к классу экспертов. Тогда эффективность предлагаемого подхода оценивалась по доле троллей, ошибочно принятых за экспертов (т.е. ошибка первого рода), а также по доле экспертов, неверно идентифицированных в качестве троллей (т.е. ошибка второго рода).

В табл. 3 приведены результаты по каждому эксперименту.

Среднее значение ошибки первого рода составило 12,50%, среднее значение ошибки второго рода составило 13,72%.

Далее представлены результаты эксперимента № 4. В табл. 4 приведена информация о пользователях и характеристики их профилей.

Количество анализируемых сообщений для каждого пользователя составило 20. Идентификаторы пользователей-троллей не указаны в связи с возможными репутационными рисками из-за публикации данных экспериментов в открытом доступе. Полученные результаты профилей пользователей были подвергнуты кластерному анализу с использованием алгоритма K-Means [17]. В результате выполнения алгоритма три пользователя-тролля включены в отдельный кластер (рис. 1).

Ошибка первого рода для данного эксперимента составила 0%, ошибка второго рода составила 50%.

Таблица 2

Результаты классификации текстовых сообщений

№	Сообщение	Истинный эмоциональный тон	Результат классификации (классификаторы 1, 2, 3, ансамбль)
1	Charlatan is EXACTLY that: misrepresenting one's skills Перевод: Шарлатан – это в точности тот, кто обесценивает чужие способности	–	–, –, –, –
2	These two leveraged-loan lifers from Eaton Vance see bright days ahead despite the negative headlines https:// Перевод: Эти два долгожителя на рынке инвестиций в кредиты из корпорации Eaton Vance смотрят в будущее с оптимизмом, несмотря на негативные заголовки в новостях	+	+, + ,–, +
3	Flashpoints of 2018: Serena Williams blows her top in the US Open final Перевод: Ключевые моменты 2018: Серена Уильямс устроила скандал в финале «Открытого чемпионата» США	+	+, +, –, +
4	I always wondered whose photos they are using for all those fake followers. Here is your answer... Перевод: Меня всегда интересовало, чьи фотографии они используют для всех этих фейковых аккаунтов своих фоловеров. А здесь ответ на этот вопрос…	–	–, + ,–, –
5	OMG. @TheOnion strikes again! Now you hack into @Bloomberg and planting satire? OMG. You guys are on fire this morning! Перевод: О боже! @TheOnion наносит ответный удар! Теперь вы взламываете @Bloomberg и распространяете сатиру по их мотивам. Ну и задали вы жару сегодня утром!	+	–, +, –, –

Таблица 3

Результаты экспериментов

Номер	Кол-во пользователей	Ошибка первого рода, %	Ошибка второго рода, %
1	15	25,00	9,09
2	15	0,00	20,00
3	10	20,00	0,00
4	9	0,00	50,00
5	16	0,00	23,08
6	13	25,00	33,33
7	12	0,00	11,11
8	15	10,00	10,00
9	14	16,67	12,05
10	15	0,00	9,09
11	15	50,00	18,18
12	10	25,00	0,00
13	10	0,00	0,00
14	10	0,00	16,67
15	11	0,00	12,50
16	12	25,00	12,50
17	14	0,00	10,00
18	11	0,00	14,29
19	14	20,00	0,00
20	9	33,33	12,50
Ср. значения	12,50	12,50	13,72

Таблица 4

Характеристики профилей поведения пользователей

Номер	Пользователь	Профиль поведения (d_1,d_2)	Соответствие признакам поведения тролля (t_1,t_2)
1	@nntaleb	(0,9557;0,8428)	(0,0)
2	@roguerad	(0,7798, 0,9714)	(0,0)
3	@HarryDCrane	(0,9303; 0,9143)	(0,0)
4	Тролль1	(0,6850; 1)	(0,1)
5	Тролль2	(0,4244; 0,2429)	(0,0)
6	Тролль3	(0; 0)	(1,1)
7	Тролль4	(0,7112; 0,7429)	(1,1)
8	Тролль5	(1; 0,6143)	(1,1)
9	Тролль6	(0,5027; 0,4571)	(1,1)

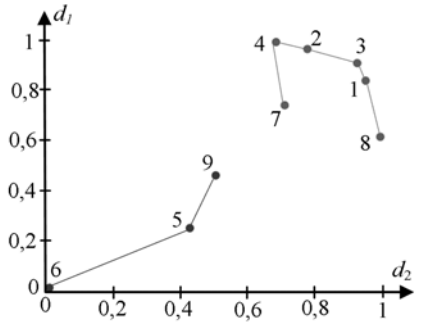

Рис. 1. Результат кластеризации поведенческих профилей

Результаты кластеризации свидетельствуют о том, что половина интернет-троллей (50%) выделена в отдельный кластер. Наличие интернет-троллей в одном кластере с экспертами объясняется способностью троллей копировать поведение обычных пользователей.

Результаты экспериментального анализа позволяют сделать вывод об эффективности ансамблевого классификатора текстовых данных для эмоционального тона сообщений социальной сети Twitter. Предложенный подход построения и оценки профилей пользователей позволяет идентифицировать явных интернет-троллей. При этом можно отметить низкие средние значения ошибок первого и второго рода, что составляет преимущество подхода.

Заключение

Представленный в статье подход к созданию профилей поведения обеспечивает повышение эффективности идентификации пользователей-экспертов за счет исключения интернет-троллей и экспертов с предвзятым мнением из выборки пользователей. Достигнутые показатели точности классификации подтверждают практическую значимость разработанного классификатора. Однако в некоторых случаях ошибки первого и второго рода достигают 50%, что характерно для выборок данных с сообщениями троллей, копирующих поведение обычных пользователей. Данные значения ошибок могут быть сокращены за счет повышения точности классификации текстовых данных посредством использования механизмов внимания, которые позволят сконцентрировать анализ эмоционального тона на отдельных частях сообщения, а также за счет расширения набора характеристик профиля пользователя, отличающих действия эксперта от действий копирующего его поведение тролля.

Работа выполнена при финансовой поддержке РФФИ (проект №17-07-00034 А).

Литература

1. Сравнение способов идентификации пользователей социальных сетей, являющихся экспертами в заданной предметной области / Е.Е. Лунева, А.А. Ефремов, Е.А. Кочегурова, П.И. Банокин, В.С. Замятина // Системы управления и информационные технологии. – 2017. – № 4(70). – С. 63–68.

2. What is an internet troll? // The Guardian [Электронный ресурс]. – Режим доступа: URL: https://www.theguardian.com/technology/2012/jun/12/what-is-an-internet-troll, свободный (дата обращения: 01.12.2018).

3. Exposing Paid Opinion Manipulation Trolls / T. Mihaylov, I. Koychev, G. Georgiev, P. Nakov // Proceedings of the International Conference Recent Advances in Natural Language Processing. – Hissar, Bulgaria: INCOMA Ltd. Shoumen, 2015. – P. 443–450.

4. Chu T. Comment abuse classification with deep learning. / T. Chu, K. Jue, M. Wang // Stanford University Электронный ресурс]. – Режим доступа: URL: https://web.stanford.edu/class/cs224n/reports/2762092.pdf, свободный (дата обращения: 01.12.2018).

5. Dollberg S. The Metadata Troll Detector: semester work. – Zurich, Switzerland, 2015. – 18 p. [Электронный ресурс]. – Режим доступа: URL: https://pub.tik.ee.ethz.ch/students/2014-HS/SA-2014-32.pdf, свободный (дата обращения: 01.12.2018).

6. Kim Y. Convolutional Neural Networks for Sentence Classification // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. – Strouds-

burg, USA: Association for Computational Linguistics. – 2014. – P. 1746–1752.

7. Банокин П.И., Ефремов А.А., Лунева Е.Е., Кочегурова Е.А. Исследование применимости рекуррентных сетей lstm в задаче поиска пользователей-экспертов социальных сетей // Программные системы и вычислительные методы. – 2017. – № 4. – С. 53–60.

8. Hong J. Sentiment analysis with deeply learned distributed representations of variable length texts / J. Hong, M. Fang. Технический отчет. – Stanford, USA: Stanford University, 2015. – 9 p. [Электронный ресурс]. – Режим доступа: URL: https://cs224d.stanford.edu/reports/Hong James.pdf, свободный (дата обращения: 01.12.2018).

9. Roy A., Kapil P., Basak K., Ekbal A. An Ensemble approach for Aggression Identification in English and Hindi Text // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC–2018). – Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. – P. 66–73.

10. Yu L.-C., Wang J., Lai K.R., Zhang X. Refining word embeddings for sentiment analysis // Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Stroudsburg, USA: ACL, 2017. – P. 545–550.

11. Wang Y., Huang M., Zhu X., Zhao L. Attention-based LSTM for aspect-level sentiment classification // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. – Austin, Texas: ACL, 2016. – P. 606–615.

12. Лунева Е.Е., Ефремов А.А., Банокин П.И. Способ идентификации пользователей-экспертов в социальных сетях // Программные системы и вычислительные методы. – 2018. – № 4. – С. 86–101.

13. Luneva E.E., Zamyatina V.S., Banokin P.I., Yefremov A.A. Estimation of social network user's influence in a given area of expertise // Journal of Physics: Conference Series. – 2017. – Vol. 803, № 1. – P. 1–6.

14. Shannon C.E. A Mathematical Theory of Communication // Bell System Technical Journal. – 1948. – №27. – P. 379–423.

15. Koizumi R. Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens? // Vocabulary Learning and Instruction. – 2012. – №1. – P. 60–69.

16. Pennington J., Socher R., Manning C.D. Pennington J. GloVe: Global Vectors for Word Representation // Standfor NLP [Электронный ресурс]. – Режим доступа: URL: https://nlp.stanford.edu/projects/glove/, свободный (дата обращения: 01.12.2018).

17. Hartigan J.A. Algorithm AS 136: A K-Means Clustering Algorithm / J.A. Hartigan, M.A. Wong // Journal of the Royal Statistical Society, Series C (Applied Statistics). – 1979. – № 28. – P. 100–108.

Банокин Павел Иванович
Ассистент инженерной школы
информационных технологий и робототехники (ИШИТР)
Национального исследовательского
Томского политехнического университета (НИ ТПУ)
Ленина пр-т, д. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 60-63-86
Эл. почта: pavel805@gmail.com

Лунева Елена Евгеньевна
Доцент ИШИТР НИ ТПУ
Ленина пр-т, д. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 60-63-86
Эл. почта: lee@tpu.ru

Ефремов Александр Александрович
Ст. преп. ИШИТР НИ ТПУ
Ленина пр-т, д. 2, г. Томск, Россия, 634050
Тел.: +7 (382-2) 60-63-86
Эл. почта: alexyefremov@tpu.ru

Banokin P.I., Luneva E.E., Yefremov A.A.

Classification of Twitter social network expert users

The behavior of social network trolls and the necessity of influential users' behavior evaluation are considered. The process of behavior profiles creation and analysis is presented. Behavior profiles are created by analysis of text messages and stored as real-value vectors. Ensemble classifier is used for evaluation of text data sentiment. Results of behavior profiles clustering are shown and discussed.

Keywords: ensemble classifier, expert user, behavior profile, internet troll, social network.
doi: 10.21293/1818-0442-2019-22-2-61-66

References

1. Luneva E.E., Yefremov A.A., Kochegurova E.A., Banokin P.I., Zamyatina V.S. The comparison of identification methods of social network users regarded as subject-matter experts. *Control Systems and Information Technology*, 2017, no. 4(70), pp. 63–68 (In Russ.).

2. What is an internet troll? *The Guardian*. Available at: https://www.theguardian.com/technology/2012/jun/12/what-is-an-internet-troll (Accessed: December 01, 2018).

3. Mihaylov T., Koychev I., Georgiev G., Nakov P. Exposing Paid Opinion Manipulation Trolls. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2015, pp. 443–450.

4. Theodora Chu, Kylie Jue, and Max Wang. Comment abuse classification with deep learning. *Stanford University*. Available at: https://web.stanford.edu/class/cs224n/reports/2762092.pdf (Accessed: December 01, 2018).

5. Dollberg S. *The Metadata Troll Detector: semester work*, Swiss Federal University of Technology Zurich, 2015, 18 p. Available at: https://pub.tik.ee.ethz.ch/students/2014-HS/SA-2014-32.pdf (Accessed: December 01, 2018).

6. Kim Y. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1752.

7. Banokin P.I., Yefremov A.A., Luneva E.E., Kochegurova E.A. Study of the applicability of LSTM recurrent networks in the task of finding users expert social networks // *Software systems and computational method*, 2017, no. 4, pp. 53–60 (in Russ.).

8. Hong J., Fang M. *Sentiment analysis with deeply learned distributed representations of variable length texts*. Stanford University, 2015, 9 p. Available at: https://cs224d.stanford.edu/reports/HongJames.pdf, свободный (Accessed: December 01, 2018).

9. Roy A., Kapil P, Basak K., Ekbal A. An Ensemble approach for Aggression Identification in English and Hindi Text. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 2018, pp. 66–73.

10. Yu L.-C., Wang J., Lai K.R., Zhang X. *Refining word embeddings for sentiment analysis*. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017, pp. 545–550.

11. Wang Y., Huang M., Zhu X., Zhao L., *Attention-based LSTM for aspect-level sentiment classification*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615.

12. Luneva E.E., Efremov A.A., Banokin P.I. A method of social network expert users identification. *Software systems and computational methods*, 2018, no. 4, pp. 86–101 (In Russ.).

13. Luneva E.E., Zamyatina V.S., Banokin P.I., Yefremov A.A. *Estimation of social network user's influence in a given area of expertise*. Journal of Physics: Conference Series, 2017, vol. 803, no. 1, pp. 1–6.

14. Shannon C.E. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948, no. 27, pp. 379–423.

15. Koizumi R. Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens? *Vocabulary Learning and Instruction*, 2012, no. 1, pp. 60–69.

16. Pennington J., Socher R., Manning C.D. GloVe: Global Vectors for Word Representation. *Stanford NLP*. Available at: https://nlp.stanford.edu/projects/glove/ (Accessed: December 01, 2018).

17. Hartigan J.A.; Wong M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 1979, no. 28, pp. 100–108.

Pavel I. Banokin
Assistant Professor,
School of Information Systems and Robotics,
National Research Tomsk Polytechnic University
2, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 60-63-86
Email: pavel805@gmail.com

Elena E. Luneva
Candidate of Engineering, Associate Professor,
School of Information Systems and Robotics,
National Research Tomsk Polytechnic University
2, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 60-63-86
Email: lee@tpu.ru

Alexander A. Yefremov
Senior Lecturer,
School of Information Systems and Robotics,
National Research Tomsk Polytechnic University
2, Lenin pr., Tomsk, Russia, 634050
ORCID 0000-0001-8149-3641
Phone: +7 (382-2) 60-63-86
Email: alexyefremov@tpu.ru