УДК 81-114.2

Семина Т. А.
Московский государственный областной университет
141014, Московская обл., г. Мытищи, ул. Веры Волошиной, д. 24, Российская Федерация

ИЗВЛЕЧЕНИЕ МНЕНИЯ АВТОРА ЧЕРЕЗ ОБРАТНУЮ ЧАСТОТУ ДОКУМЕНТА

АННОТАЦИЯ

Целью исследования выступает выявление закономерностей для автоматического поиска мнения автора в статьях, относящихся к жанру информационной журналистики, который накладывает ограничение на выражение оценочных суждений. Автором собран корпус политических статей, для лексем из этой коллекции рассчитана мера из сферы информационного поиска: обратная частота документа. Эмпирически показана применимость этого критерия для поиска сниженной лексики, которая является индикатором наличия мнения автора. В заключение сформулированы направления дальнейшего исследования этой проблемы и рассмотрен вопрос выявления мнения автора через обратную частоту документа в других текстах.

КЛЮЧЕВЫЕ СЛОВА:
анализ тональности, информационный поиск, обратная частота документа, частота термина, мнение.

СТРУКТУРА

T. Semina
Moscow Region State University
24 Very Voloshinoy ul., Mytyshi 141014, Moscow reg., Russian Federation

AUTHOR'S OPINION MINING USING INVERSE DOCUMENT FREQUENCY

ABSTRACT

The purpose of the article is to identify patterns for author's opinion mining in articles relating to the genre of information journalism, that imposes a restriction on the expression of judgments and opinions. A corpus of political articles has been assembled by the author and the metric from information

retrieval called inverse document frequency has been counted for every token in the collection. The applicability of this criterion has been shown for the search of colloquial language which is an indicator of the presence of the author's opinion. In conclusion, the directions of further research of this problem are formulated and the question of identifying the author's opinion through the inverse document frequency in other texts is considered.

KEY WORDS:
sentiment analysis, information retrieval, inverse document frequency, term frequency, opinion

ВВЕДЕНИЕ

Анализ тональности, известный как сентимент-анализ или система извлечения мнений, – это область изучения мнений, оценок и эмоций людей по отношению к таким объектам, как продукты, организации, личности, события, проблемы, и их атрибутам. Сентимент (от англ. "sentiment" – «чувство, мнение, настроение») – эмоциональная оценка, выраженная в тексте, также называемая тональностью текста [5, с. 511].

Мнение, как считает большинство исследователей в области анализа тональности, состоит из главных компонентов: субъекта (или источника), объекта (или цели) и тональности [10, p. 2505]. Субъектом называется человек, которому принадлежит мнение, в ряде случаев субъектом может быть не человек, а компания или организация. Объект определяется как предмет, человек или атрибут, о котором высказано мнение. Тональность – это полярность мнения, она может быть бинарной, т. е. положительной и отрицательной, возможно добавление нейтральной тональности или градуированной оценки [6, с. 39]. В современном анализе тональности понятие «мнение» сильно отличается от словарного значения этого слова. Мнение в анализе тональности не обязательно содержит оценку или эмоциональную окраску, при анализе на некотором материале мнение считается действием субъекта, которое положительно или отрицательно влияет на объект [8, p. 472].

МЕТОДЫ

Исследования тональности относятся как к теоретической лингвистике, так и к компьютерной лингвистике. Теория языка рассматривает вербальные средства выражения мнения и ряд сопутствующих задач, например проблему интерпретации сарказма, структуру отношений между сущностями, характеристики фрагментов текста, определяющих авторскую позицию. Анализ тональности как раздел компьютерной лингвистики рассматривает те же вопросы, но ставит целью не поиск закономерностей и создание теорий, а практическое решение задач. Стоит отметить, что эти два взгляда на тональность дополняют друг друга. Системы анализа тональности на основе

машинного обучения меньше зависят от особенностей материала исследования и могут не иметь в своей основе результатов теоретического анализа, тем не менее системы на основе правил (или инженерные системы) невозможно построить без первоначального анализа эксперта-лингвиста.

Несмотря на то, что одна из первых работ по анализу тональности (сам термин появился позднее, в то время это направление называлось анализом субъективности) имела в качестве материала тексты художественной литературы, в настоящее время для его проведения используют только тексты из интернета [11]. Это связано с доступностью огромных массивов данных, потенциально содержащих мнение автора текста: кинорецензии, отзывы на различные товары или услуги, тексты Twitter'а, блоги и статьи. В связи с этим можно отменить необходимость апробации данных интернет-лингвистики [2]. Тексты блогов и микроблогов отличаются от обычного письменного модуса дискурса, поэтому построение качественной практической модели невозможно без понимания особенностей интернет-текстов.

В данном исследовании в качестве материала рассматриваются аналитические политические статьи, в которых упоминается Россия. Статьи взяты с сайта *inosmi.ru*, на котором размещаются переводы публикаций зарубежных журналов. Подобные тексты имеют значительное отличие от текстов отзывов и блогов: автор текста не может открыто высказывать своё мнение об обсуждаемых событиях, это запрещается в жанрах информационной журналистики[1]. Тем не менее встречались случаи, когда автор текста давал свою оценку персонам или событиям, но делал это через иронию [3, с. 165]. Подобные высказывания отличаются от остального текста, создавая желаемый автором статьи эффект: читатель видит отношение автора к обсуждаемым в статье событиям, но при этом языковых единиц, прямо выражающих мнение, нет.

Как правило, при исследовании новостных текстов с точки зрения анализа тональности учёные рассматривают мнения именованных сущностей по отношению друг к другу [9, p. 1324]. Именованная сущность (или просто сущность) – объект, человек или событие, имеющее чётко определённого референта, соответственно, некоторые типы именованных сущностей, такие как персоны или геополитические образования, могут являться субъектом или источником мнения, все типы сущностей являются объектом мнения [4].

При рассмотрении аналитических статей можно увидеть, что ряд тональных отношений имеет в качестве субъекта не сущность, а автора текста, что, как было написано ранее, не подразумевается в жанрах информационной журналистики. Тем не менее автор текста не выражает мнения напрямую, ис-

[1] См.: Колесниченко А. В. Настольная книга журналиста: учебное пособие. М.: Аспект Пресс, 2013. 334 с.

пользуя эксплицитные оценочные конструкции, вместо этого применяются ирония, сарказм и стилистически сниженная лексика. Примером подобной лексики будет выражение «альфа-самец» по отношению к Президенту РФ В. В. Путину:

(1) *Британское решение покинуть ЕС вполне устраивает российского* ***альфа-самца****, — такова оценка старшего научного сотрудника Датского института международных исследований и специалиста по России Флеминга Сплидбёля (Flemming Splidsboel)*[2]**.**

Показанный в примере (1) способ выражения мнения затрудняет автоматический поиск подобных тональных отношений, потому что при любом подходе к анализу необходимо вносить дополнительные правила для поиска таких оценок. При машинном обучении в выборку вся подобная лексика не попадёт, следовательно, на практике система будет пропускать эту информацию. Добавить список этих лексем в систему на основе правил тоже не представляется возможным, так как списков таких выражений не существует.

Для поиска подобной ироничной лексики в данном исследовании предлагается использовать обратную частоту документа (idf – inverse document frequency) [7, p. 246], так как её показатель для редких слов в корпусе будет выше, чем для частотных. Метрика idf обычно применяется в информационном поиске и машинном обучении для нормализации векторного пространства, но для каждого термина в документе она применяется с частотой термина в документе (tf – term frequency). Произведение tf*idf (часто записывается tf-idf) позволяет оценить степень важности каждой языковой единицы для описания документа. Наша задача состоит не в описании терминов каждого документа, а в поиске ироничных терминов во всей коллекции, поэтому частота термина tf была убрана из расчётов.

$idf_t = \log (N / df_t)$

В анализе тональности tf-idf применяется в тех случаях, когда он рассматривается как задача классификации, тогда эта метрика помогает «взвесить» термины и определить их значимость. Существует и так называемая delta-схема взвешивания признаков, она учитывает не распространение признака во всём корпусе текстов, а неравномерность распределения слова в классах тональности. Чем более неравномерно распределено это слово или выражение, тем выше будет его вес при delta-схеме [1, с. 157].

Простой поиск слов с низкой частотой в языке не дал бы результатов, потому что стилистически сниженная лексика может, наоборот, быть частотной в речи. Метрика idf позволит найти слова, которые редко встречаются не в языке, а в отобранных текстах.

[2] Линдегорд С. Путин считает brexit победой, 2016 // ИноСМИ.Ру. URL: https://inosmi.ru/politic/20160626/236981189.html (дата обращения: 02.02.2019).

В настоящее время используемый нами корпус содержит 151 статью общим объёмом 158905 слов. Из анализа исключены стоп-слова (предлоги, союзы, местоимения) и именованные сущности. Максимальный показатель idf при подобных данных будет равен 2.1790.

if df = 1:

idf = 2.1790

Как было сказано ранее, частота термина (term frequency) не учитывалась в расчётах, тем не менее для некоторых слов с высокой обратной частотой документа эта метрика проверялась, почти во всех случаях она была равна 1, т. е. это слово встретилось всего один раз в корпусе.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Рассмотрим выдержки из статей с сайта *inosmi.ru*, в которых содержатся выражения с высокими показателями idf. После каждого слова указана округлённая обратная частота документа, в ряде случаев считалась idf и для выражений, таких как *вести себя* или *друг с другом*.

(2) *И всё же обе* 0.9749 *стороны* 0.0821 *ведут* 0.5992 *себя* 0.3216 [как коллокация – 1.2247] *сейчас* 0.3594, *как молодые* 1.1790 ***влюбленные* 2.1790**. *В конце* 0.5069 *июня* 0.4800 *Эрдоган и Путин 40* 1.4008 *минут* 1.4800 *говорили* 0.5557 [друг с другом] 1.0328 *по* ***телефону* 1.8779** — *впервые* 0.9485 *за последние* 0.2813 *полгода* 1.4800. *До этого Кремль* ***демонстративно* 1.8779** *не отвечал* 0.6227 *на телефонные* 1.3339 ***звонки* 1.8779** *из Анкары*[3]**.**

В примерах (2) – (4) выделены лексемы с idf ≥ 1.5769 как слова с низкой частотой в корпусе. На данный момент до конца не определён минимальный показатель idf для лексики, которая может выражать мнение автора, для представления первых результатов исследования было принято решение установить число 1.5769 как минимальное для редкой лексики. В примере (2) в число выделенных выражений вошли слова *влюбленные, телефон, демонстративно, звонки*. При подсчёте коэффициентов учитывалась нормальная форма слова, а не словоформа. Слова *влюбленные* и *демонстративно* показывают отношение автора к В. В. Путину, Р. Т. Эрдогану и возглавляемым ими странам, данные лексемы в политической статье выглядят инородно и не относятся к жанру дискурса статьи. Выражения *телефон* и *звонки* не являются яркими индикаторами мнения автора, однако их уместность в рамках информационного жанра журналистики кажется спорной. Необходимо уточнить, что при анализе статей стоит цель найти мнение сущности о другой сущности, поэтому дублирующиеся тональные отношения не учитываются, следовательно, если *телефон* и *звонки* относятся к группе ложноположительных элементов, это не снизит качество поиска мнений.

[3] Биддер Б. Примирение Москвы с Турцией: как молодые влюбленные // ИноСМИ.Ру. URL: https://inosmi.ru/politic/20160707/237102541.html (дата обращения: 02.02.2019).

(3) *Иногда* 0.9749 *он* [В. В. Путин] {ведёт себя} 1.2247 *как* ***капризный*** **2.1790** ***ребенок*** **1.5769**, *который пытается* 0.5355 *добиться* 0.6108 *от* ***родителей*** **2.1790** *своего, бросая* 1.1376 ***игрушки*** **2.1790**. *Иногда* 0.9749 *как политический* 0.1496 ***хулиган*** **1.5769**. *А иногда* 0.9749 *он напоминает* 0.9237 *Хрущева,* ***стучащего*** **1.8779** ***дырявым*** **1.8779** ***ботинком*** **2.1790** ***по трибуне*** **1.5769**. *Но сейчас* 0.3594 *он оказался* 0.2705 *в тупике* 1.1376: *Запад не ведется* 1.4800 *на эти игры* 1.0029[4]**.**

В примере (3) ряд лексем, которые могли бы считаться маркерами мнения автора, не превысили установленного нами показателя idf ≥ 1.5769, что является одной из причин возможного снижения этого показателя в будущем. Возможно, в число сниженной лексики стоит включать лексемы с idf > 1, но проверить правильность установленной границы можно только после проведения анализа текстов и серии экспериментов. В примере (3) ряд лексем имеет высокую обратную частоту документа, что указывает на отношение автора к В. В. Путину. Слово дырявый в тексте имело tf = 2, т. е. в статье слово было употреблено дважды, но второй раз в контексте дырявая память. В примере встречаются не просто единичные элементы с высокой idf, но целые группы таких слов: капризный ребенок, стучащего дырявым ботинком по трибуне. Эти примеры показывают, что можно искать не одиночные лексемы, а кластеры лексем. Ограничения в объёме корпуса не позволяют гарантировать, что при подобном анализе в число маркеров мнения автора не будет включена лексика, которая могла редко встречаться в корпусе, при этом не являться оценочной, в то время как поиск словосочетаний с элементами, имеющими высокую idf, может снизить потенциальное количество ошибочных результатов поиска.

(4) *Между тем резервы* 1.2759 *эти и так с* ***гулькин*** **2.1790** ***нос*** **1.7019**, *а перспектива* 0.5258 *резкой* 0.6349 ***девальвации*** **1.5769** *рубля* 1.0998 *после* ***деноминации*** **2.1790**, *призванной* 0.9237 *укрепить* 0.3936 *к нему доверие* 0.7988, *– это для* ***белорусского*** **1.7019** *руководства* 0.5355 *просто* ***ужас*** **1.7019,** ***летящий*** **2.1790** *на* ***крыльях*** **2.1790** *ночи* 1.4800. *Так что для спасения* 1.4800 *от обвала* 1.4008 *очередной* 0.8568 *конфликт* 0.3728 *с Москвой надо бы* ***уладить*** **2.1790** *как можно скорее* 0.6349. *Однако* 0.2058 *Кобяков очевидно* 0.7640 *не получал* 0.3216 *от Лукашенко* ***распоряжения*** **1.8779** ***ехать*** **2.1790** *в Москву с белым* **1.8779** ***флагом*** **1.5769.** ***Капитуляция*** **1.7019** *в этом вопросе* 0.1456 *может лишь* ***раззадорить*** **2.1790** *Кремль, станут* 0.0998 ***дожимать*** **2.1790** *и в другом* 0.0998[5]**.**

В примере (4) встретились кластеры слов с высокой idf: *гулькин нос, ужас, летящий на крыльях ночи, распоряжения ехать*. Встречаются и одиноч-

[4] Велиньский Б. Коморовский изменил взгляд Обамы // ИноСМИ.Ру. URL: https://inosmi.ru/military/20160711/237148612.html (дата обращения: 03.02.2019).

[5] Класковский А. Москва будет выбивать у Лукашенко газовый долг // ИноСМИ.Ру.URL: https://inosmi.ru/economic/20160714/237186452.html (дата обращения: 03.02.2019).

ные слова *раззадорить, дожимать*. *Деноминация* и *девальвация* относятся к ложноположительным результатам поиска, т. е. наш алгоритм ошибочно выделил бы их как маркеры мнения автора. Наличие таких лексем в корпусе, скорее, является недочётом при сборе статей, так как все остальные тексты относятся к политике и не затрагивают вопросов экономики. Для анализа тональности большое значение имеет монотематичность текстовой коллекции, поэтому подобную статью можно убрать из используемого корпуса. Слово *белорусского* имеет высокую обратную частоту документа по схожей причине, оно не относится к основной тематике остальных текстов, кроме того, во время сбора статей для анализа не было большого количества публикаций об отношениях между Россией и Белоруссией, это могло стать причиной отсутствия других словоформ этой лексемы в коллекции.

ВЫВОДЫ

В заключение стоит отметить, что изучение распределения обратной частоты документа находится на начальном этапе. В ходе дальнейшего исследования будет продолжено рассмотрение этого явления в аналитических политических статьях. Исходя из результатов первого исследования, можно сделать ряд выводов, имеющих большое значение для дальнейшей работы над этой проблемой.

Итак, выводы. Во-первых, автоматическое извлечение мнения автора в текстах, не позволяющих эксплицитно выражать свою позицию, является возможным при применении предложенного в статье алгоритма, имеющего в основе использование метрики из информационного поиска. Во-вторых, проблема определения минимального показателя idf для выявления сниженной лексики ещё не решена и требует дополнительного анализа данных. В-третьих, этот метод не применим на текстах блогов или рецензий, потому что орфографические ошибки затруднили бы поиск и выдавали бы много ложноположительных результатов. В-четвёртых, при установлении мнения, субъектом которого является автор, стоит в первую очередь учитывать кластеры лексем с высокой обратной частотой документа. Кроме того, возможно рассмотрение количества подобных слов в определённом отрезке текста, чтобы избежать появления ложноположительных результатов из-за несоответствия темы статьи или её фрагмента остальной коллекции.

ЛИТЕРАТУРА

1. Алексеев А. А., Кугуракова В. В., Иванов Д. С. Выявление психологического портрета на основе определения тональности сообщений для антропоморфного социального агента // Электронные библиотеки. 2016. Т. 19. № 3. С. 149–165.
2. Ахренова Н. А. Интернет-лингвистика: новая парадигма в описании языка интернета // Вестник Московского государственного областного университета. Серия: Лингвистика. 2016. № 3. С. 8–14.

3. Гейко Н. Р., Сиривля М. А. Имплицитные оценки в политическом дискурсе // Вестник Брянского государственного университета. 2016. № 2 (28). С. 164–166.
4. Можарова В. А., Лукашевич Н. В. Исследование признаков для извлечения именованных сущностей из текстов на русском языке // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2017. № 5. С. 14–21.
5. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог», Бекасово, 25–29 мая 2011 г. Вып. 10 (17). М.: Издательство РГГУ, 2011. С. 510–522.
6. Семина Т. А. Дихотомия субъективность vs. объективность и тональная релевантность в задачах анализа тональности // Вестник Московского государственного областного университета. Серия: Лингвистика. 2018. № 1. С. 38–45.
7. Chen K., Zhang Z., Long J., Zhang H. Turning from TF-IDF to TF-IGM for term weighting in text classification // Expert Systems With Applications. 2016. № 66. P. 245–260.
8. Choi Y., Wiebe J., Mihalcea R. Coarse-grained +/– Effect Word Sense Disambiguation for Implicit Sentiment Analysis // The IEEE Transactions on Affective Computing. 2017. Vol. 8. № 4. P. 471–479.
9. Deng L., Wiebe J. MPQA 3.0: An Entity/Event-Level Sentiment Corpus // Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. Denver, Colorado, 2015. P. 1323–1328.
10. Peng M., Zhang Q., Jiang Y. G. Cross-Domain Sentiment Classification with Target Domain Specific Information // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018. P. 2505–2513.
11. Wiebe J. M. Tracking Point of View in Narrative // Computational Linguistics. 1994. Vol. 20. № 2. P. 233–287.

REFERENCES

1. Alekseev A. A., Kugurakova V. V., Ivanov D. S. [The identification of a psychological portrait based on the definition of key messages for anthropomorphic social agent]. In: *Elektronnye biblioteki* [Electronic Libraries], 2016, vol. 19, no. 3, pp. 149–165.
2. Akhrenova N. A. [Internet linguistics: a new paradigm in the description of the language of the Internet]. In: *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika* [Bulletin of the Moscow Region State University. Series: Linguistics], 2016, no. 3, pp. 8–14.
3. Geiko N. R., Sirivlya M. A. [Implicit evaluations in political discourse]. In: *Vestnik Bryanskogo gosudarstvennogo universiteta* [Bulletin of Bryansk State University], 2016, no. 2 (28), pp. 164–166.
4. Mozharova V. A., Lukashevich N. V. [Examination of the indications for the extraction of named entities from texts in Russian]. In: *Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnye protsessy i sistemy* [Scientific and technical information. Series 2: Information Processes and Systems], 2017, no. 5, pp. 14–21.

5. Pazel'skaya A. G., Solov'ev A. N. [The method of definition of emotions in Russian texts]. In: *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog», Bekasovo, 25–29 maya 2011 g. Vyp. 10 (17)* [Computational linguistics and intellectual technologies: based on the materials of the annual International Conference "Dialogue", Bekasovo, May 25–29, 2011. Iss. 10 (17)]. Moscow, Publishing house of the Russian State University for the Humanities Publ., 2011, pp. 510–522.
6. Semina T. A. [The dichotomy of subjectivity vs. objectivity and tone the relevance in the task of sentiment analysis]. In: *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika* [Bulletin of the Moscow Region State University. Series: Linguistics], 2018, no. 1, pp. 38–45.
7. Chen K., Zhang Z., Long J., Zhang H. Turning from TF-IDF to TF-IGM for term weighting in text classification. In: *Expert Systems With Applications*, 2016, no. 66, pp. 245–260.
8. Choi Y., Wiebe J., Mihalcea R. Coarse-grained +/– Effect Word Sense Disambiguation for Implicit Sentiment Analysis. In: *The IEEE Transactions on Affective Computing*, 2017, vol. 8, no. 4, pp. 471–479.
9. Deng L., Wiebe J. MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In: *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. Denver, Colorado, 2015, pp. 1323–1328.
10. Peng M., Zhang Q., Jiang Y. G. Cross-Domain Sentiment Classification with Target Domain Specific Information. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, 2018, pp. 2505–2513.
11. Wiebe J. M. Tracking Point of View in Narrative. In: *Computational Linguistics*, 1994, vol. 20, no. 2, pp. 233–287.

ДАТА ПУБЛИКАЦИИ

Статья поступила в редакцию: 21.03.2019

Статья размещена на сайте: 04.06.2019

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Семина Татьяна Алексеевна – ассистент кафедры теоретической и прикладной лингвистики Московского государственного областного университета; e-mail: taniasemina@gmail.com

Tat'yana A. Semina – assistant lecturer at the Department of Theoretical and Applied Linguistics, Moscow Region State University; e-mail: taniasemina@gmail.com

ПРАВИЛЬНАЯ ССЫЛКА НА СТАТЬЮ / FOR CITATION

Семина Т. А. Извлечение мнения автора через обратную частоту документа // Вестник Московского государственного областного университета (электронный журнал). 2019. № 2. URL: www.evestnik-mgou.ru

Semina T. A. Author's opinion mining using inverse document frequency. In: *Bulletin of Moscow Region State University (e-journal)*, 2019, no. 2. Available at: www.evestnik-mgou.ru