ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2021 Управление, вычислительная техника и информатика № 54

УДК 519.2
DOI: 10.17223/19988605/54/7

Д.К. Левоневский, А.И. Савельев

ПОДХОД И АРХИТЕКТУРА ДЛЯ СИСТЕМАТИЗАЦИИ И ВЫЯВЛЕНИЯ ПРИЗНАКОВ АГРЕССИИ В РУССКОЯЗЫЧНОМ ТЕКСТОВОМ КОНТЕНТЕ

Работа выполнена при финансовой поддержке фонда РФФИ № 18-29-22061_МК.

Рассматриваются признаки агрессии в русскоязычных текстах, выполняется классификация этих признаков. Предлагаются способы автоматизации выявления признаков и для обработки естественного языка и программных средств общего назначения. Разработана архитектура программной системы, выполняющей векторизацию текстовых сообщений. Реализуемый в этой архитектуре подход позволяет оценивать признаки агрессивности текстового контента с достаточной точностью, а погрешность обусловлена в основном многозначностью слов.

Ключевые слова: обработка естественного языка; анализ тональности; извлечение эмоций; агрессия; анализ текста.

Компьютерные сети накапливают большое количество разнородной информации, анализ которой востребован, но трудно реализуем. Одной из востребованных задач является обнаружение проявлений агрессии в сетевом контенте. В данной работе рассматриваются вопросы анализа текстового контента. Сложная структура текстового контента требует уменьшения его размерности для применения методов анализа. Для классификации текстовых сообщений, применения методов машинного обучения необходимо выполнять векторизацию текста на основе некоторых признаков. Эта задача рассматривается в данной статье.

Значительное количество методов классификации текстов и, в частности, выявления агрессивных сообщений рассмотрено в статьях [1, 2]. Как правило, они реализуют типовую последовательность операций обработки данных, включающую удаление неинформативных компонентов текста, токенизацию, разметку частей речи и используют для извлечения признаков подходы bag-of-words, bag-of-stems и т.п. [3]. Реализация таких подходов возможна и для русскоязычного контента, но одним из необходимых условий их эффективности являются качественные размеченные наборы данных [4]. Другой сложностью является многозначность слов и иной характер корреляции между их негативной семантикой и агрессивностью сообщения в целом. В [5] исследуется характер языковых различий при использовании устойчивых выражений. В статье [6] с этим связываются существенно более низкие возможности выявления агрессивных сообщений на русском языке.

В [7] рассматриваются психолингвистические и контентные признаки агрессивного поведения в чатах. В этой статье анализируется англоязычный контент, однако большая часть признаков и принципов их построения применима и для других языков, в частности ряд категорий лексических признаков, признаки, связанные с частями речи (прежде всего с глаголами и местоимениями). Кроме того, авторы анализируют поведение пользователей в чате.

В работе [8] выделены следующие лексические и дискурсивные языковые средства, использующиеся при агрессивном речевом поведении. К первой группе относятся средства языка, выражающие негативную оценку (жаргонная лексика, окказиональные слова, инвективная лексика и пр.). Ко второй группе относятся дискурсивные средства, формирующиеся непосредственно в тексте в процессе общения и, как правило, выражающие скрытую речевую агрессию (языковая демагогия, интертекстуальность и пр.). Перечисленные языковые средства неравнозначны, их сложно формализовать. Кроме того,

для интернет-пространства характерны использование смеси текстового и аудиовизуального контента, смеси разных языков, специфической лексики, наличие сообщений с ошибками, опечатками, нетипичной пунктуацией, что существенно затрудняет применение лексического анализа текстов.

В работе [9] обсуждаются вопросы, связанные с автоматическим выявлением в текстах социальных сетей проявлений вербальной агрессивности. Показано, что при автоматическом анализе эмоциональной насыщенности текста представляется целесообразным использовать слова, выражающие негативную оценку. Авторы работы разработали систему психолингвистического анализа текстовой информации (PLATIn). Данная система основана на использовании словарей русского текста, разбитых на списки лексических единиц, согласно их семантической направленности. К недостаткам данного метода следует отнести то, что он не делает различий на семантико-синтаксическом уровне.

В [10, 11] рассматриваются психолингвистические, лексические, семантические маркеры, которые можно использовать для характеристики агрессивности текста. К психолингвистическим маркерам относятся количество слов в предложении, коэффициент определенности действия, количество глаголов в пассивном залоге, средняя длина слова, отношение количества инфинитивов к общему числу глаголов и т.д. Эти характеристики могут выявить эмоциональное состояние автора, наличие призывов к действию, противопоставлений «мы–они» и т.п. К словарным лексическим маркерам относятся: обозначение негативных эмоциональных и телесных состояний (гнев, отвратительный), слова с деструктивной семантикой (уничтожить, раздавить), лексика физического насилия (бить, ранить), инвективная лексика (идиот, предатель) и т.д. Семантические маркеры выражают значения слов, например: деструктив – объект разрушающего воздействия (взорвать дом), ликвидатив – объект, прекращающий существование (убить человека), результатив – следствие (привести к кризису) и т.д.

Таким образом, источники приводят множество признаков агрессии в тексте, но при этом нет их общей классификации с точки зрения природы этих признаков и, следовательно, подходов к их выявлению. Кроме того, последовательности операций обработки данных зависят от конкретных признаков и должны учитывать их особенности.

1. Классификация признаков

Рассмотренные признаки можно сгруппировать в следующие категории:

1. Лексические: к этим признакам относятся отдельные слова и устойчивые выражения, сигнализирующие о возможности агрессивного контента.

2. Морфологические: признаки, связанные с частями слов и словообразованием, образованием неологизмов и производных слов с помощью суффиксов, приставок.

3. Статистические: признаки, основанные на частотах использования частей речи, знаков препинания.

4. Дискурсивные: самая сложная в выявлении группа, связанная с использованием демагогии, различной стилистики речи, иронии и сарказма, искажений слов и других приемов, которые сложно формализовать и выявить.

5. Косвенные: маркеры эмоциональной выраженности, связанные в основном с формой представления текста (маскировка, регистр и т.п.).

Примеры приведенных признаков представлены в табл. 1.

Таблица 1

Признаки агрессии в тексте

Класс средств / признаков	Примеры
Лексические признаки	
Инвективная, обсценная, стилистически сниженная лексика	Подлец, двурушник, враг народа
Эвфемизмы	Женщина с низкой социальной ответственностью
Жаргонная лексика	Фуфлыжник
Немотивированное использование иноязычных элементов в целях агрессивного воздействия на читателя, провоцирующее возникновение у него чувства неполноценности из-за непонятности изложения	Да Вы просто рутинёр, милейший! Ваши слова, уважаемый, бурлеск чистой воды. Ровно как и Вы – акциденция современности

Окончание табл. 1

Класс средств / признаков	Примеры
Глаголы с общим значением разрушительного действия	Грызть, дырявить, бить
Глаголы уничтожения	Зарезать, губить, пепелить
Глаголы повреждения	Ранить, царапать, ковырять
Слова с негативной семантикой	Украсть, хапнуть
Названия профессий, употребляемые в переносном значении	Палач, мясник
Зоосемантические метафоры	Кобель, кобыла, свинья
Морфологические признаки	
Окказициональные слова – неологизмы, оценочность которых может быть связана как с мотивирующей (производящей) базой (например, собственные имена лиц, слова с негативным денотатом и т.д.), так и со словообразовательными средствами (например, маркированные суффиксы -щина, -ость, -ация, -изм, -ист, -ец, -оид, размерно-оценочные суффиксы, префиксы а-, без-, анти-, контр-, де-, квази-, псевдо-, экс- и т.д.)	Троцкизм, обломовщина, либероид, ватник, коммуняки, дерьмократы, прихватизация
Дискурсивные признаки	
Языковая демагогия (сознательное нарушение словесных пресуппозиций, постулатов успешного общения, использование речевых импликатур)	Ну Вы у нас признанный эксперт, конечно
Тенденциозное использование негативной информации, перегруженность текста негативной информацией (например, использование безысключительной лексики: каждый, все, никто и т.д.)	Все знают, что ты отсталый
Интертекстуальность: обращение к вербальным прецедентным феноменам, которые связаны с определенными эмоциями и оценками для людей, разделяющих знание о них, с целью иронизирования, насмешки и т.д.	Шариков
Псевдоимперативы	Поговори мне еще!
Риторические вопросы	Ты первый день на работе, что ли?
Гиперболы	Миллион раз я тебе говорил
Статистические признаки	
Количество местоимений 1-го, 3-го лица множественного числа	Управы на них нет! Да мы их в пыль сотрём
Относительная доля глаголов	Будем рвать их, резать, надо прибить к чертям
Относительная доля глаголов в будущем времени	Разберемся еще, время покажет
Большое количество восклицательных знаков, кратные восклицательные и вопросительные знаки	Твари!!!!!
Косвенные признаки	
Маскировка слов (замена букв спецсимволами, цифрами, буквами других алфавитов, намеренное искажение слов)	п0д0нки
Маркеры эмоциональной выраженности: использование верхнего регистра, наличие опечаток, ошибок	ТЫ ЧТО ТВОРИШ???!!!

Следует отметить, что наличие этих признаков не означает, что текстовое сообщение агрессивно, а только характеризует повышенную вероятность этого.

2. Выявление признаков

Среди языков программирования одним из наиболее предпочтительных для решения задачи является Python в силу его широких возможностей с точки зрения обработки текстов и наличия гибких структур данных. Поэтому все упоминаемые библиотеки приводятся так, как они названы в Python.

Для выявления лексических признаков используются словари. Наиболее эффективным видится применение в качестве основного словаря русского раздела Wiktionary, так как он обладает следующими преимуществами:

– имеет регулярно публикуемые дампы в форматах SQL и XML для компьютерной обработки;

– содержит не только начальные формы слов, но и падежные формы, формы множественного числа и др.;

– содержит как слова общей лексики, так и большое количество бранных и жаргонных слов;

– содержит много устойчивых выражений;

– имеет списки синонимов и антонимов, что позволяет отслеживать семантические связи между понятиями;

– предусматривает большое количество категорий слов, характеризующихся пометками «грубое», «уничижительное», «вульгарное», «бранное», «криминальный жаргон», «сниженное», «пренебрежительное», «просторечное» и др.

Однако надо отметить, что словарь часто не содержит новейших форм интернет-жаргона и производных форм слов, что является частью более общей проблемы неполноты данных в задаче обнаружения агрессивного контента [12].

Выявление морфологических признаков может выполняться с помощью библиотек разбора слов по составу (XMorphy) или регулярных выражений (re).

Статистические признаки имеют различные способы выявления:

– для признаков, основанных на употреблении частей речи, используются технические средства классификации слов (например, библиотека pymorphy2, позволяющая определить часть речи, узнать категории времени, лица, числа);

– для признаков, основанных на пунктуации, смайликов можно использовать простые средства поиска, например регулярные выражения (библиотека re);

– для поиска ошибок можно использовать средства автоматической проверки орфографии (библиотека enchant).

Дискурсивные признаки сложнее всего поддаются анализу, но в ряде случаев их можно свести к признакам других типов [13].

Надо отметить, что большинство методов выполняет классификацию сообщений, реализуя примерно одну и ту же последовательность операций по обработке текстов, которая включает удаление стоп-слов, токенизацию, разметку частей речи, стемминг и т.д., а также типовые способы извлечения признаков (например, построение векторов на основе «мешка слов»). Предлагаемый подход в целом соответствует такой схеме, но отличается конструированием различных последовательностей блоков преобразования данных для разных типов признаков, после чего выполняется объединение результатов. Соответствующая диаграмма классов показана на рис 1.

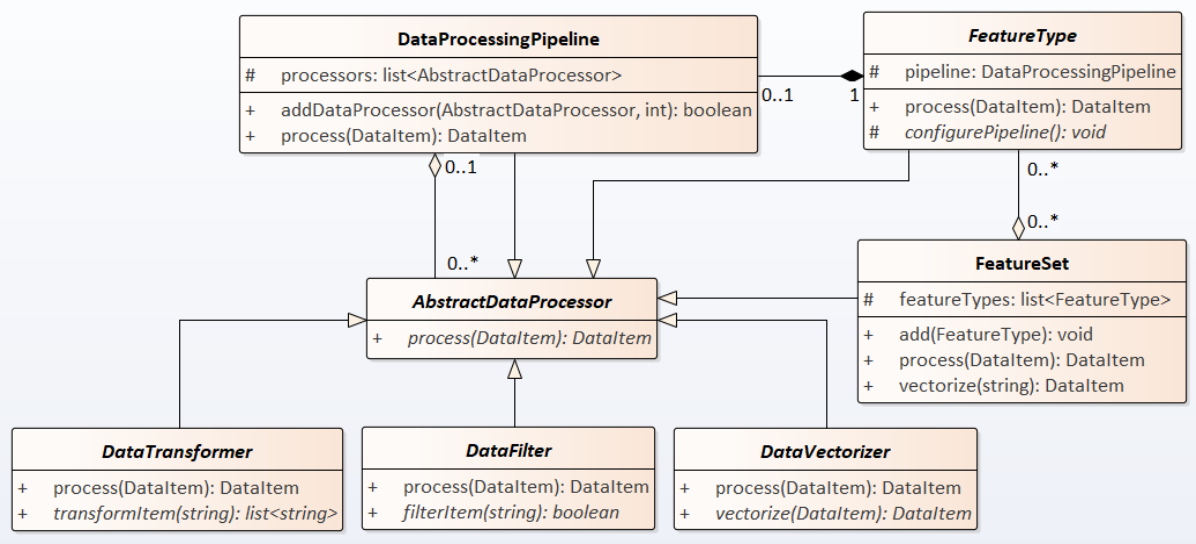


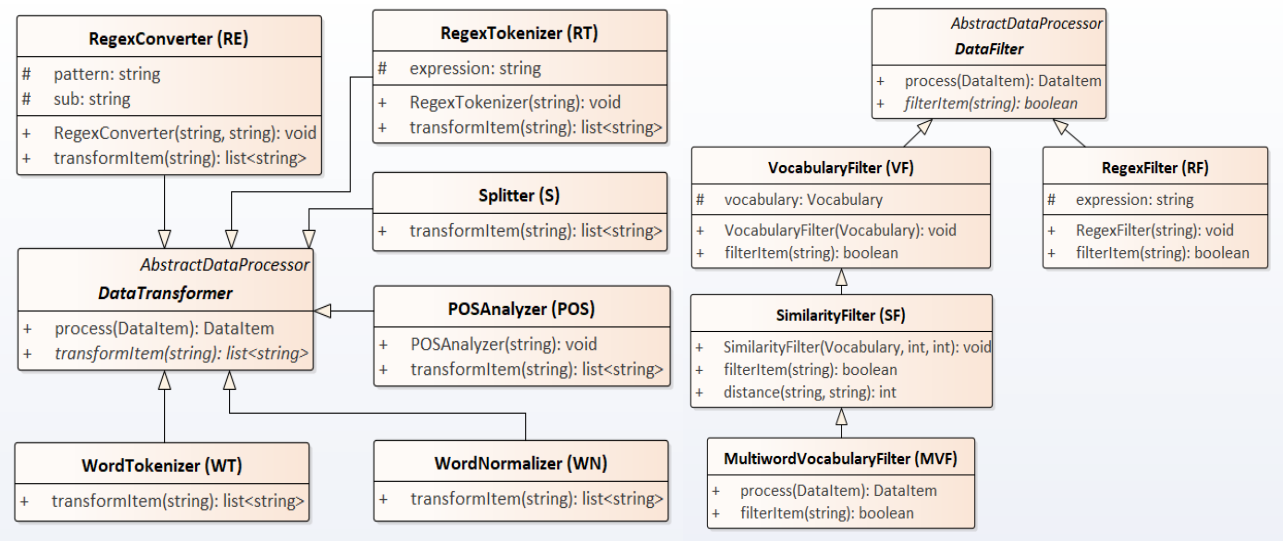
Рис. 1. Общая диаграмма классов для построения набора признаков
Fig. 1. High-level class diagram for building the feature set

В приведенной диаграмме полный набор признаков описывается объектом FeatureSet, содержащим отдельные признаки – объекты класса FeatureType. Каждый признак имеет свою последовательность обработки данных (DataProcessingPipeline), состоящую из последовательных блоков пре-

образования данных (объекты классов-наследников AbstractDataProcessor). Сами классы FeatureSet, FeatureType и DataProcessingPipeline также наследуют AbstractDataProcessor, выполняя роль фасада и позволяя одним и тем же способом выполнять как атомарные, так и составные преобразования. Основные виды классов обработки данных – DataProcessor (изменение формы представления данных), DataFilter (фильтрация) и DataVectorizer (преобразование в число или вектор). Их подклассы отображены на рис. 2.

Преобразование данных (рис. 2, *a*) выполняется с использованием библиотек работы с текстом (POSAnalyzer, POS; WordNormalizer, WS) и регулярных выражений (WordTokenizer, WT; RegexConverter, RE; RegexTokenizer, RT; Splitter, S).

Фильтрация данных (рис. 2, *b*) производится на основе словарей (VocabularyFilter, VF) и регулярных выражений (RegexFilter, RF). При нестрогой словарной фильтрации (SimilarityFilter, SF) используется расстояние Левенштейна [14], рассчитываемое как минимальное количество односимвольных операций вставки, удаления или замены, необходимых для преобразования одного слова в другое. Эта величина выбрана, так как она отражает редакционное расстояние между словами и характеризует сходство слов.



a *b*

Рис. 2. Классы: *a* – преобразования данных, *b* – фильтрации данных
Fig. 2. Classes for *a* – data transformation, *b* – data filtration

Таблица 2

Последовательность обработки данных для распознавания признаков агрессии

Признаки	Последовательность блоков преобразования данных
Отдельные лексические единицы – инвективная лексика; стилистически сниженная лексика; жаргон; зоосемантические метафоры; отдельные немногочисленные семантические группы слов (например, маркеры всеобщности)	WT → WN → SF → IC
Устойчивые выражения, метафоры	WT → MVF → IC
Морфологические признаки (приставки, суффиксы)	WT → WN → RF → IC
Глаголы в повелительном наклонении	WT → POS → VF → RC
Количество местоимений 1-го, 3-го лица множественного числа	WT → WN → VF → IC
Относительная доля глаголов	WT → POS → VF → RC
Относительная доля глаголов в будущем времени	WT → POS → VF → RC
Кратные знаки препинания	RE → RT → RF → IC
Наличие опечаток, ошибок	WT → WN → VF → IC
Использование верхнего регистра	S → RF → RF → RC
Маскировка слов	RT → RF → IC

Окончательное значение признака рассчитывается как количество найденных образцов в тексте (ItemCounter, IC) или относительная частота их появления или отсутствия (RateCounter, RC).

Схема обработки данных при распознавании различных типов признаков приведена в табл. 2. Для каждого признака она включает набор блоков обработки данных, которые применяются последовательно в порядке, указанном символом «→».

3. Результаты

Для тестирования был использован dataset [15], содержащий комментарии с русскоязычных сайтов, в большинстве случаев агрессивного характера. Для оценки точности часть данных из dataset в размере n = 300 элементов была размечена вручную. Затем для каждого признака f были рассчитаны значения ошибки при i-м измерении E_{fi}, среднее значение ошибки $\overline{E}_f$, его среднеквадратическое отклонение σ_f и оценка верхнего порога ошибки, рассчитанная по правилу 3σ в предположении, что распределение величины подчиняется нормальному закону:

$$E_{fi} = \left| f_i^a - f_i^m \right|,$$

где f_i^a – значение признака, определенное автоматически, f_i^m – то же значение, определенное вручную;

$$\overline{E}_f = \frac{1}{n}\sum_{i=1}^{n} E_{fi};$$

$$\sigma_f = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_{fi} - \overline{E}_f)^2}.$$

Результаты оценивания приведены в табл. 3.

Таблица 3

Результаты оценивания точности расчета признаков

Признак	Среднее значение ошибки $\overline{E}_f$	Среднеквадратическое отклонение σ_f	Пороговое значение
Количество лексических единиц	1,9	2,1	8,4
Количество устойчивых выражений	0,08	0,37	1,2
Относительная доля использования верхнего регистра	0	0	0
Количество случаев использования кратных знаки препинания	0,01	0,1	0,3
Относительная доля глаголов	0,03	0,05	0,18
Относительная доля глаголов в повелительном наклонении	0,002	0,01	0,04
Относительная доля глаголов в будущем времени	0,004	0,01	0,04
Количество местоимений 1-го, 3-го лица множественного числа	0	0	0
Морфологические признаки	0,8	0,5	2,5
Количество опечаток, ошибок	2	1,4	6,2
Количество случаев маскировки слов	0,05	0,03	0,13

Из табл. 3 видно, что часть признаков, которые являются строго определенными и формальными, определяется практически без погрешностей. Среди остальных признаков наибольшая погрешность наблюдается при определении количества лексических единиц. То, что преобладает ошибка I рода, объясняется многозначностью слов и, в частности, тем, что слова общей лексики могут иметь дополнительные значения, несущие негативную направленность, зачастую в узкоспециализированном контексте.

Заключение

Предложенный в статье подход включает в себя классификацию признаков агрессивного текстового контента, способы их выявления и архитектуру соответствующей программной системы.

Реализация системы позволяет оценивать признаки агрессивности текстового контента с достаточной точностью. Предложенный подход может быть использован при решении более общей задачи классификации многомодального контента, включающего также речь и видео [16, 17]. Ограничения подхода состоят в том, что его нельзя применить для сообщений, содержащих разнородный контент или смесь языков.

ЛИТЕРАТУРА

1. Mäntylä M.V., Graziotin D., Kuutila M. The evolution of sentiment analysis – a review of research topics, venues, and top cited papers // Computer Science Review. 2018. V. 27. P. 16–32.
2. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey // Ain Shams Engineering Journal. 2014. V. 5, № 4. P. 1093–1113.
3. Ventirozos F.K., Varlamis I., Tsatsaronis G. Detecting aggressive behavior in discussion threads using text mining // International Conference on Computational Linguistics and Intelligent Text Processing. Springer, Cham, 2017. P. 420–431.
4. Levonevskiy D., Malov D., Vatamaniuk I. Estimating Aggressiveness of Russian Texts by Means of Machine Learning // International Conference on Speech and Computer. Springer, Cham, 2019. P. 270–279.
5. Zykova I.V. Perception of verbal communication reflected in Russian and English phraseology: towards a new theory of Phraseologism-formation // Procedia – social and behavioral sciences. 2016. V. 236. P. 139–145.
6. Gordeev D. Automatic detection of verbal aggression for Russian and American imageboards // Procedia - Social and Behavioral Sciences. 2016. V. 236. P. 71–75.
7. Parapar J., Losada D.E., Barreiro A. Combining Psycho-linguistic, Content-based and Chat-based Features to Detect Predation in Chatrooms // J. UCS. 2014. V. 20, № 2. P. 213–239.
8. Петрова Н.Е., Рацибурская Л.В. Язык современных СМИ: Средства речевой агрессии. М. : Флинта : Наука, 2011. 160 с.
9. Девяткин Д.А., Кузнецова Ю.М., Чудова Н.В., Швец А.В. Интеллектуальный анализ проявлений вербальной агрессивности в текстах сетевых сообществ // Искусственный интеллект и принятие решений. 2014. № 2. С. 27–41.
10. Ковалёв А.К., Кузнецова Ю.М., Минин А.Н., Пенкина М.Ю., Смирнов И.В., Станкевич М.А., Чудова Н.В. Методы выявления по тексту психологических характеристик автора (на примере агрессивности) // Вопросы кибербезопасности. 2019. Т. 4, № 32. С. 72–79.
11. Сбоев А.Г., Гудовских Д.В., Молошников И.А., Кукин К.А., Рыбка Р.Б., Иванов И.И., Власов Д.С. Автоматическое выделение психолингвистических характеристик текстов в рамках концепции Big Data // Современные информационные технологии и ИТ-образование. 2013. № 9. С. 433–438.
12. Reyes A., Rosso P. Making objective decisions from subjective data: Detecting irony in customer reviews // Decision support systems. 2012. V. 53, № 4. P. 754–760.
13. Rosa H., Pereira N., Ribeiro R., Ferreira P.C., Carvalho J.P., Oliveira S., Coheur L., Paulino P., Veiga Simão A.M., Trancoso I. Automatic cyberbullying detection: a systematic review // Computers in Human Behavior. 2019. V. 93. P. 333–345.
14. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. 1965. Т. 163, № 4. С. 845–848.
15. Russian Language Toxic Comments Small dataset with labeled comments from 2ch.hk and pikabu.ru. URL: https://www.kaggle.com/blackmoon/russian-language-toxic-comments (accessed: 04.08.2020).
16. Уздяев М.Ю., Левоневский Д.К., Шумская О.О., Летенков М.А. Метод детектирования агрессивных пользователей информационного пространства на основе генеративно-состязательных нейронных сетей // Информационно-измерительные и управляющие системы. 2019. Т. 17, № 5. С. 60–68.
17. Уздяев М.Ю. Распознавание агрессивных действий с использованием нейросетевых архитектур 3D-CNN // Известия ТулГУ. Технические науки. 2020. № 2. С. 316–330.

Поступила в редакцию 5 августа 2020 г.

Levonevskiy D.K., Saveliev A.I. (2021) APPROACH AND ARCHITECTURE FOR CATEGORIZATION AND REVEAL OF AGGRESSION FEATURES IN RUSSIAN TEXT CONTENT. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science] 54. pp. 56–64

DOI: 10.17223/19988605/54/7

Methods and systems, involved in detection of aggressive features in text content, find sufficiently broad application. Particularly, they are used for analysis of comments and reviews online, sentiment analysis, concerning certain events, development of digital assistants for moderators of network discussions, etc. Thereby, complex structure of the text content requires reduction of text dimensions for analysis implementation. This paper discusses aggression features in Russian text. The relevant sources consider various text-based aggression features, but lack an overarching classification of these features, that would be built on their common foundations. Complex structure of textual content requires dimension reduction in it to apply analytic methods. For classification of textual

messages and employment of machine learning methods, text vectorization is necessary, based on certain features. From this perspective, such a classification is established in this paper. All the features are divided into five classes: lexical, morphological, statistical, conversational and indirect. Specific words and set expressions, which can possibly denote violence, are classified as the aggressive content. Morphological features have to do with morphems and word building, neologism creation and derivative words using suffixes or prefixes. Statistical features have to do with the frequency of certain parts of speech and punctuation. Discource features is the set of features most difficult to be extracted, because these features have to do with demagogy, different stylistic variations of speech, sarcasm and irony, word distortion and other approaches, which are difficult to formalize and reveal. Indirect features are the markers of emotional expressiveness, having to do mostly with the presentational aspects of the text (masking, letter case, etc.). Features and items for each class are aggregated in the tabular view. Approaches are proposed to automate these feature detection processes; these approaches are based on thesauri, natural language processing libraries and generic software tooling. The architecture of a software suite is developed for text message vectorization. It is specified with class diagrams, describing the functional units of relevant transformations, filtering and vectorization of text content. The approach, implemented in this architecture, generally fits into the framework, which includes stop-word removal, tokenization, part-of-speech tagging, stemming and other common approaches to feature extraction. The specificity of this approach consists in establishment of different sequences of data transformations for different kinds of features, then result aggregation is performed. Recognition accuracy estimation is also performed. The approach, proposed in this paper, allows to estimate the aggression features in the text content with decent accuracy, whereas the estimation errors arise primarily because of polysemy, particularly, because common words can have additional negative meanings, relevant for certain niche contexts. Limitations of this approach preclude to use it for analysis of heterogeneous content or mix of languages or mix of languages.

Keywords: natural language processing; sentiment analysis; emotion mining; aggression; text analysis.

LEVONEVSKIY Dmitriy Konstantinovich (Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation).
E-mail: DLewonewski.8781@gmail.com

SAVELIEV Anton Igorevich (Candidate of Engineering Sciences, Senior Researcher St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation).
E-mail: saveliev.ais@yandex.ru

REFERENCES

1. Mäntylä, M.V., Graziotin, D. & Kuutila, M. (2018) The evolution of sentiment analysis – A review of research topics, venues, and top cited papers. *Computer Science Review*. 27. pp. 16–32. DOI: 10.1016/j.cosrev.2017.10.002
2. Medhat, W., Hassan, A. & Korashy, H. (2014) Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. 5(4). pp. 1093–1113. DOI: 10.1016/j.asej.2014.04.011
3. Ventirozos, F.K., Varlamis, I. & Tsatsaronis, G. (2017) Detecting aggressive behavior in discussion threads using text mining. *International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, Cham. pp. 420–431.
4. Levonevskiy, D., Malov, D. & Vatamaniuk, I. (2019) Estimating Aggressiveness of Russian Texts by Means of Machine Learning. *International Conference on Speech and Computer*. Springer, Cham. pp. 270–279.
5. Zykova, I.V. (2016) Perception of verbal communication reflected in Russian and English phraseology: towards a new theory of Phraseologism-formation. *Procedia–Social and Behavioral Sciences*. 236. pp. 139–145. DOI: 10.1016/j.sbspro.2016.12.052
6. Gordeev, D. (2016) Automatic detection of verbal aggression for Russian and American imageboards. *Procedia-Social and Behavioral Sciences*. 236. pp. 71–75. DOI: 10.1007/978-3-319-11581-8_40
7. Parapar, J., Losada, D.E. & Barreiro, A. (2014) Combining Psycho-linguistic, Content-based and Chat-based Features to Detect Predation in Chatrooms. *Journal of UCS*. 20(2). pp. 213–239. DOI: 10.3217/jucs-020-02-0213
8. Petrova, N.E. & Ratsiburskaya, L.V. (2011) *Yazyk sovremennykh SMI: Sredstva rechevoy agressii* [The Language of Modern Media: Means of Verbal Aggression]. Moscow: Flinta: Nauka.
9. Devyatkin, D.A., Kuznetsova, Yu.M., Chudova, N.V. & Shvets, A.V. (2014) An intellectual analysis for the social web: recognizing verbal aggression. *Iskusstvennyy intellekt i prinyatie resheniy – Artificial Intelligence and Decision Making*. 2. pp. 27–41.
10. Kovalev, A.K., Kuznetsova, Yu.M., Minin, A.N., Penkina, M.Yu., Smirnov, I.V., Stankevich, M.A. & Chudova, N.V. (2019) Text analysis approach for identifying psychological characteristics (with aggressiveness as an example). *Voprosy kiberbezopasnosti – Cybersecurity Issues*. 4(32). pp. 72–79. DOI: 10.21681/2311-3456-2019-4-72-79
11. Sboev, A.G., Gudovskikh, D.V., Moloshnikov, I.A., Kukin, K.A., Rybka, R.B., Ivanov, I.I. & Vlasov, D.S. (2013) Avtomaticheskoe vydelenie psikholingvisticheskikh kharakteristik tekstov v ramkakh kontseptsii Big Data [Automatic highlighting of the psycholinguistic characteristics of texts within the Big Data concept]. *Sovremennye informatsionnye tekhnologii i IT-obrazovanie – Modern Information Technology and IT-education*. 9. pp. 433–438.
12. Reyes, A. & Rosso, P. (2012) Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*. 53(4). pp. 754–760.
13. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P.C., Carvalho, J.P., Oliveira, S., Coheur, L., Paulino, P., Veiga Simão, A.M. & Trancoso, I. (2019) Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*. 93. pp. 333–345. DOI: 10.1016/j.chb.2018.12.021

14. Levenshtein, V.I. (1965) Dvoichnye kody s ispravleniem vypadeniy, vstavok i zameshcheniy simvolov [Binary codes capable of correcting deletions, insertions, and reversals]. *Doklady Akademii Nauk SSSR*. 163(4). pp. 845–848.
15. Kaggle.com. (n.d.) *Russian Language Toxic Comments Small dataset with labeled comments from 2ch.hk and pikabu.ru*. [Online] Available from: https://www.kaggle.com/blackmoon/russian-language-toxic-comments (Accessed: 1st August 2020).
16. Uzdiaev, M.Yu., Levonevsky, D.K., Shumskaya, O.O. & Letenkov, M.A. (2019) Methods for detecting aggressive users of the information space based on generative-competitive neural networks. *Informatsionno-izmeritel'nye i upravlyayushchie sistemy – Information-Measuring and Control Systems*. 17(5). pp. 60–68. DOI: 10.18127/j20700814-201905-08
17. Uzdiaev, M.Yu. (2020) Violent Action Recognition Using 3D CNN Neural Network Architectures. *Izvestiya TulGU. Tekhnicheskie nauki*. 2. pp. 316–330.