Анализ тональности текста с использованием методов машинного обучения

Sentiment Analysis of Text Using Machine Learning Techniques

А.С. Романов [1]
Aleksandr Romanov [1]
alexx.romanov@gmail.com

М.И. Васильева [1]
Maria Vasilieva [1]
maryska-98@mail.ru

А.В. Куртукова [1]
Anna Kurtukova [1]
av.kurtukova@gmail.com

Р.В. Мещеряков [1]
Roman Meshcheryakov [1]
mrv@ieee.org

[1] ФГОУ ВО Томский государственный университет систем управления и радиоэлектроники
Томск, Российская Федерация

[1]Tomsk State University of Control Systems and Radioelectronics
Tomsk, Russian Federation

Abstract

The paper deals with the sentiment analysis of text and provides an results of experiments with classification texts into positive, negative and neutral classes using different machine learning techniques (support vector machine, naive Bayes classifier and Random Forest) combining with manually prepared vocabularies. In addition, review of research, methods and software products for sentiment analysis are given.

Keywords: *sentiment analysis, support vector machine, social networks*

Аннотация

В статье приводятся результаты исследования методики анализа тональности текста с использованием методов машинного обучения, таких как метод опорных векторов, наивный Байесовский классификатор, методы случайных деревьев. Приводится обзор исследований, методов и программных продуктов в области анализа

тональности текста, описываются этапы моделирование процесса проведения экспериментов и определения тональности текста, приводятся описания созданных корпусов текстов и словарей, а также полученные результаты исследований.

Ключевые слова: *тональность текста, метод опорных векторов, социальные сети*

1 Введение

Создание комплексной автоматизированной системы интеллектуального анализа текстовой информации, способной, в частности, определять автора и пол автора текста, возраст, уровень образования, эмоциональное состояние автора в момент написания текста, анализировать авторство исходных текстов программ для различных языков программирования, идентифицировать искусственно созданные тексты, находить фрагменты заимствований и плагиат и т.д. является актуальной задачей. Настоящая статья просвещена исследованию проблемы определения тональности текста. Под тональностью будем понимать эмоционально окрашенную лексику и эмоциональную оценку, выраженную автором относительно чего-либо. Анализ тональности имеет важное практическое применение.

1. Оценка качества товаров и услуг на основе отзывов пользователей Интернет-ресурсов. Объектом эмоционального оценивания может быть имя собственное, название продукта, организации, услуги, профессии и проч., по отношению к которому выражается мнение. Однако ежедневное количество публикуемых отзывов в социальных сетях достигает огромного количества, поэтому обработка отзывов вручную оказывается невозможной и требует автоматизации.

2. Противодействие экстремизму и терроризму. Мониторинг социальных сетей и методы автоматизированного анализа текста используются силовыми структурами для фильтрации и выявления сообщений, содержащих информацию о противоправных действиях, готовящихся террористических атаках, контента, содержащего запрещенный материалы и т.д.

3. Автороведческое исследование документов в криминалистике. Анализ тональности позволяет дополнить модель автора текста, в частности определить эмоциональное состояние автора в момент написания текста, точнее идентифицировать пол автора и т.д.

4. Анализ ситуации на фондовых рынках и прогнозирование волатильности финансовых активов. На основе анализа тональности текстов новостных лент, обзоров финансовых аналитиков, отчетов трейдеров, речей высокопоставленных чиновников и руководителей, а также общего настроения пользователей социальных сетей определяется корреляция этих данных с трендами фондовых рынков и строятся прогнозы изменения цен финансовых активов.

5. Составление текстов с заранее заданными эмоциональными характеристиками. На основе словарей позитивных и негативных слов, а также методов анализа тональности можно составлять тексты, несущие определенное эмоциональное воздействие на потенциальную аудиторию. Активно применяется в репутационном маркетинге, в сфере связей с общественностью, при написании речей политических деятелей, а также в образовательных технологиях для создания легкоусвояемых учебных материалов.

Очевидно, что совершенствование имеющихся методов анализа тональности является актуальной задачей. Но стоит отметить, что проблема активно исследуется отече-

ственными и зарубежными учеными.

В работе [Kotelnikov 2012] авторы ставят перед собой цель протестировать и сравнить, во-первых, различные подходы к представлению текста в рамках векторной модели, во-вторых, различные методы машинного обучения, а именно метод опорных векторов, наивный байесовский классификатор, метод ключевых слов и его комбинацию с методом опорных векторов. В качестве исходных данных использовался корпус отзывов пользователей по трём группам товаров (цифровые фотокамеры, книги и фильмы). Данный корпус был размечен экспертами РОМИП. В работе для всех текстов использовалась предобработка. В качестве векторной модели текста авторы предлагают использовать бинарную модель с косинусной нормализацией, т.к. она наиболее выгодна с точки зрения эффективности и вычислительной сложности. Аналогичный подход применялся в [Pang 2002]. При разделении на два класса лучше всего себя показал метод опорных векторов с точностью равной 95%. Комбинированный метод на основе метода опорных векторов и ключевых слов при разделении на три класса показал точность 79%. При пяти классах точность комбинированного метода составила 49%.

В [Yusupova 2012] в качестве алгоритма машинного обучения был выбран наивный байесовский классификатор. Для повышения точности классификации использовались методы построения ансамбля классификаторов (бустинг и бэггинг). Признаковое пространство рассматривается в виде “мешка слов”. Для обучения и оценки точности классификации использовался тестовый набор, состоящий из отзывов клиентов российских банков. Он включает в себя 304 положительных и 850 негативных отзывов на русском языке. Лучший результат (точность 87,69%) удалось достичь при использовании бэггинга на основе мультиномиальной модели наивного байесовского классификатора.

В статье [Sida 2012] авторы получили аналогичные выводы для английского языка – модель Бернулли оказалась хуже, чем мультиномиальная. Единственный случай, где они сравнимы – это при использовании униграмм в качестве классифицирующего признака. С короткими сообщениями лучше справилась мультиномиальная модель (точность 83,55%), а с длинным текстом – метод опорных векторов (89,16%). Комбинированный метод на основе наивного байесовского классификатора и метода опорных векторов справился хорошо как с длинными текстами (91,22%), так и с короткими сообщениями (89,45%). В работе [Aksenov 2012] текст анализируется двумя разными алгоритмами: методом максимальной энтропии и методом опорных векторов. Для того чтобы избежать проблемы переобучения, алгоритмы тестируются методом перекрестной проверки. Наиболее эффективным оказался метод максимальной энтропии, достигнувший точности классификации 81,5% при полноте 0,79. Метод опорных векторов показал точность 82% при полноте 0,75.

В работе [Lukashkina 2015] автор, в качестве исходных данных использует два англоязычных корпуса: первый, состоящий из 2000 отзывов по 1000 на каждый класс, и второй, содержащий отзывы покупателей на различные товары, объемом 6000 отзывов. В работе использовалась предобработка данных и скользящий контроль с разбиением выборки на 8 частей. В качестве классификаторов использовались: метод опорных векторов с RBF ядром, метод опорных векторов с линейным ядром, наивный байесовский классификатор и двухслойная нейронная сеть с сигмоидальной функцией активации. Также в работе использовалась композиция метода опорных векторов и наивного байесовского классификатора и принятие решения голосованием по большинству. При этом лучшие результаты были получены при композиции алгоритмов, также было показано, что точность классификации существенно зависит от тематики отзыва.

Также активно ведется разработка программных систем, автоматизирующих процесс оценки тональности текста. Такие программы как "Аналитический курьер" [Analytical Courier], "RCO Fact Extractor SDK" [RCO], "ВААЛ" [VAAL], "Eureka Engine" Eureka имеют богатый функционал. Однако существует нехватка решений для анализа текста на русском языке. Кроме того, качественный инструмент должен учитывать специфику области применения (например, особенности сообщений в социальной сети «Твиттер») и проводить классификацию не только на положительный и отрицательный классы, но и нейтральный.

2 Технология исследования

На рис. 1 представлена методика проведения экспериментов в нотации IDEF0. Входами являются размеченный корпус данных, состоящий из множества пар «пример-класс» и словари для предобработки текста. Результатом проведения экспериментов являются отчет о классификации, который содержит необходимую информацию для определения качества классификации и обученный классификатор для анализа тональности текста. Обученный классификатор может использоваться в дальнейшем, минуя процесс обучения.

Задача классификации текстовой информации определяется следующим образом. Пусть существует описание документа $d \in X$, где X - векторное пространство документов, и фиксированный набор классов $C = \{c_1, c_2..., c_m\}$. Из обучающей выборки (множества документов с заранее известными классами) $D = \{\langle d, c\rangle \mid \langle d, c\rangle \in X \times C\}$ с помощью метода обучения G необходимо получить классифицирующую функцию $G(D) = \gamma$, которая отображает документы в классы $\gamma : X \to C$.

Упоминавшийся в работах выше классификатор на основе машины опорных векторов показал отличные результаты при решении авторами данной статьи смежных задач, связанных с интеллектуальной обработкой текста в работах [Romanov 2009], [Romanov 2010], [Romanov 2011], [Sozinova 2014], [Romanov 2014], поэтому был выбран как основной. Также исследовались методы Naive Bayes и Random Forest как наиболее часто встречающиеся инструменты принятия решений в работах, связанных с обработкой текста.

Все элементы из обучающей и тестовой выборки представляют собой n-мерные векторы признаков. В задачах обработки текста на естественном языке очень распространено представление документов в виде n-грамм (последовательности элементов текста длиной n). Для $n = 1$ такая последовательность состоит из одного слова и называется униграммой. Для $n = 2$ такая последовательность называется биграммой. Для $n = 3$ такая последовательность называется триграммой. Т.о. документ определяется как вектор:

$$d = (z_1, z_2, ..., z_{|V|}),$$

где V - множество уникальных термов из обучающей выборки;

z_i - вес i терма.

В качестве способа взвешивания термов в работе применяется обратная частота документа (term frequency inverse document frequency, $tfidf$):

$$z_i = tfidf(t_i, d, D) = \frac{n_i}{\sum n_k} \times \log \frac{|D|}{|d_i \supset t_i|}$$

где $|D|$ - кол-во документов в корпусе;

$|d_i \supset t_i|$ - кол-во документов, в которых встречается t_i.

Для оценки качества классификации используется F-мера, представляющая собой гармоническое среднее между правильностью P и полнотой классификации R:

$$F = 2\frac{P \times R}{P + R}$$

Также методом оценки аналитической модели и её поведения на независимых данных является перекрестная проверка k-fold cross validation, при $k = 10$.

3 Результаты

Эксперименты проводились на нескольких размеченных корпусах.

Корпус tweets – корпус сообщений социальной сети «Твиттер», за основу которого взят корпус [Rubtsova 2012]. В настоящее время является единственным на данный момент русскоязычным корпусом, состоящим из текстов на общую тематику. Остальные корпуса представляют собой коллекции отзывов на определенную тематику, при использовании которых классификатор может давать хорошие результаты, но при смене тематики тестового набора данных качество классификации может значительно ухудшится. Сообщения из социальных сетей более эмоциональны, в то время как отзывы более продуманы и конструктивны.

Корпус содержит около 400 000 позитивных сообщений и 300 000 негативных сообщений обычных пользователей. Кроме того, в него добавлены также нейтральные сообщения общим количеством 150 000. Сбор и разметка сообщений производились с помощью специально написанного скрипта и привлеченных экспертов. В контексте задачи анализа тональности текста, основными особенностями сообщений в социальной сети Твиттер являются: малый размер сообщения (140 символов), наличие сленга, сокращений и грамматических ошибок, наличие обсценной лексики, эмотиконов и использования разговорного языка.

Корпус kinopoisk - содержит рецензии на кинофильмы и собран с Интернет-ресурса «Кинопоиск» (https://www.kinopoisk.ru). Корпус состоит из 50000 негативных рецензий, 58000 нейтральных рецензий, 380000 позитивных рецензий, общим объемом 2 Гб. Поскольку объем выборки рецензий для каждого класса варьируется, при проведении экспериментов использовалось по 50000 рецензии каждого класса.

Для учета особенностей сообщений социальной сети Твиттер потребовалось создать словари (см. табл. 1). С использованием коллекции коротких сообщений из социальной сети Твиттер и коллекции рецензий на кинофильмы, была подсчитана частота употребления для всех слов и отброшены термы с частотой менее 100. Итоговый объем составил 53750 термов. Далее итоговое множество термов было обработано вручную и распределено по словарям. Для учета контекста подсчитывались частоты употребления словосочетаний длинной в два и три слова. Словосочетание отбрасывалось если частота употребления словосочетания менее 100, отсутствовало существительное или присутствовало, но отсутствовала одна из следующих граммем: полное прилагательное, краткое прилагательное, компаратив, наречие.

Из результатов экспериментов (см. табл. 2)[1], можно сделать вывод, во всех случаях

[1]Обозначения, принятые в Таблице 2: A – точность, P – правильность, R – полнота, F1 – F-мера, t –

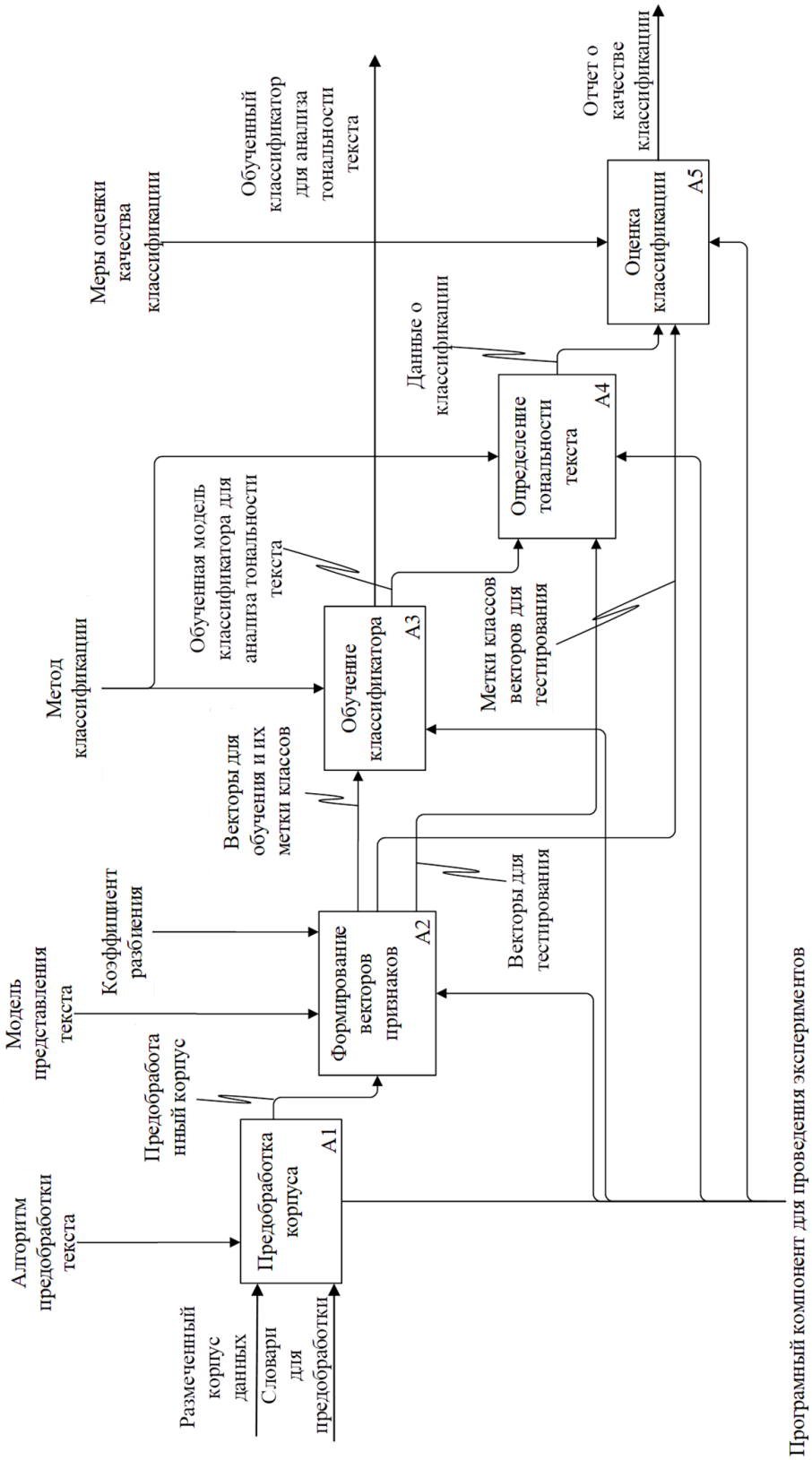

Рис. 1: Процесс проведения экспериментов

быстрее всего производит обучение и классификацию наивный байесовский классификатор, а лучшей точностью обладает классификатор на основе метода опорных векторов. Random Forest требует больше процессорного времени по сравнению с остальными классификаторами.

Таблица 1: Используемые словари

Словарь	Количество слов
Позитивно окрашенных слов	432
Негативно окрашенных слов	519
Словарь обсценной лексики	206
Словарь позитивных эмотиконов	67
Словарь негативных эмотиконов	42
Словарь стоп слов	506
Итого	1772

Также было установлено, что для русскоязычного текста применение нормализации текста и коррекции орфографии не улучшает значимо качество классификации. Не приводит к заметному улучшению результатов сокращение признакового пространства и применение различных стратегий мультиклассификации для классификатора SVM.

4 Выводы

Сравнение результатов вычислений с работами других авторов показало, что разработанная система анализа тональности текста позволяет осуществлять классификацию текстов на положительно и отрицательно окрашенные и нейтральные с точностью 98%. На сегодняшний день это один из лучших результатов.

Учитывая, что использовались привычные методы классификации, можно сделать вывод, что высокая точность обуславливается применением признаков с высокой разделяющей способностью, полученных путем тщательной ручной обработки данных.

Список литературы

[Kotelnikov 2012] Kotelnikov E.V. (2012). Avtomaticheskiy analiz tonal'nosti tekstov na osnove metodov mashinnogo obucheniya [Sentiment analysis of texts based on machine learning methods]. In Proceedings of the Conference Dialog, Vyp. 11 (18). S.7-10.(In Russian) = Котельников Е.В. Автоматический анализ тональности текстов на основе методов машинного обучения / Е.В. Котельников, М.В. Клековкина // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 11 (18). М.: Изд-во РГГУ, 2012. С. 7–10.

[Pang 2002] Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2002. Pp. 79–86.

время обучении.

[Yusupova 2012] Yusupova N.I. (2012) Algoritmicheskoye i programmnoye obespecheniye dlya analiza tonal'nosti tekstovykh soobshcheniy s ispol'zovaniyem mashinnogo obucheniya [Algorithmic and software for the analysis of the tonality of text messages using machine learning]. In Vestnik UGATU, Vol. 16, № 6 (51). S. 91-99.(In Russian) = Юсупова Н.И. Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения / Н.И. Юсупова, Д.Р Богданова, М.В Бойко // Вестник УГАТУ «Математическое моделирование, численные методы и комплексы программ», Уфа 2012 Т. 16, № 6(51). С. 91–99.

[Sida 2012] Sida W., Christopher D. (2012) Manning Stanford University. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. URL: https://goo.gl/26EDft.

[Aksenov 2012] Aksenov A.V. (2016) Analiz tonal'nosti tekstovykh soobshcheniy sotsial'noy seti Twitter [Sentiment analysis of text messages social network twitter]. In Journal Theory. Practice. Innovations. S. 4-12.(In Russian) = Аксенов А.В. Анализ тональности текстовых сообщений социальной сети Twitter. / А.В. Аксенов // Научно-технический журнал «ТЕОРИЯ. ПРАКТИКА. ИННОВАЦИИ» июль 2016. – С. 4-12.

[Lukashkina 2015] Lukashkina Yu.N (2015) Strukturnyye i statisticheskiye metody analiza emotsional'noy okraski teksta [Structural and statistical methods for analyzing the emotional coloring of a text].(In Russian) = Лукашкина Ю.Н. Структурные и статистические методы анализа эмоциональной окраски текста. URL: https://goo.gl/rFt2Zf.

[Analytical Courier] System "Analytical Courier". URL: http://www.iteco.ru/solutions/business_intelligence_products/analytical_courier

[RCO] RCO Fact Extractor SDK. URL: http://www.rco.ru/?page_id=3554.

[VAAL] VAAL project. URL: http://www.vaal.ru

[Eureka] Eureka Engine. URL: http://eurekaengine.ru/ru/description.

[Romanov 2009] Romanov A.S., Meshcheryakov R.V. (2009) Identifikatsiya avtora teksta s pomoshch'yu apparata opornykh vektorov [Authorship identification with support vector machine in case of two possible alternatives]. In Proceedings of the Conference Dialog, Vyp. 8 (15). S. 432–437.(In Russian) = Романов А.С. Идентификация автора текста с помощью аппарата опорных векторов / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). М.: РГГУ, 2009. Вып. 8 (15). С. 432–437.

[Romanov 2010] Romanov A.S., Meshcheryakov R.V. (2010) Identifikatsiya avtorstva korotkikh tekstov metodami mashinnogo obucheniya [Identification of authorship of short texts with machine learning techniques]. In Proceedings of the Conference Dialog, Vyp. 9 (16). S. 407–413.(In Russian) = Романов А.С. Идентификация авторства коротких текстов методами машинного обучения / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). М.: Изд-во РГГУ, 2010. Вып. 9 (16). С. 407–413.

[Romanov 2011] Romanov A.S., Meshcheryakov R.V. (2011) Opredeleniye pola avtora korotkogo elektronnogo soobshcheniya [Gender identification of the author of a short message]. In Proceedings of the Conference Dialog, Vyp. 10(17). S. 620–626.(In Russian) = Романов А.С. Определение пола автора короткого электронного сообщения / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25 - 29 мая 2011 г.). М.: Изд-во РГГУ, 2011. Вып. 10 (17). С. 620–626.

[Sozinova 2014] Sozinova I.S., Romanov A.S., Meshcheryakov R.V. (2014) Opredeleniye poiskovogo spama s ispol'zovaniyem metoda opornykh vektorov [Search spam identification using support vector machine]. In Proceedings of the SPIIRAN, Vyp. 36. S. 78–91.(In Russian) = Романов А.С. Определение поискового спама с использованием метода опорных векторов / И.С. Созинова, Р.В. Мещеряков // Труды СПИИРАН, 2014. Вып. 36. С. 78–91.

[Romanov 2014] Romanov A.S., Rezanova Z.I., Meshcheryakov R.V. (2014) Metodika proverki odnorodnosti teksta i vyyavleniya plagiata na osnove metoda opornykh vektorov i fil'tra bystroy korrelyatsii [Plagiarism detection and text homogeneity checking technique based on one-class support machine and fast correlation-based filter]. In Reports of TUSUR, № 2 (32). S. 264–269.(In Russian) = Романов А.С. Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции / З.И. Резанова, Р.В. Мещеряков // Доклады ТУСУР. № 2 (32). Томск: Издательство ТУСУР, 2014. С. 264–269.

[Rubtsova 2012] Rubtsova Yu. (2012) Avtomaticheskoye postroyeniye i analiz korpusa korotkikh tekstov (postov mikroblogov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora [Automatic construction and analysis of the body of short texts (microblogging posts) for the task of developing and training a tone classifier]. In Inzheneriya znaniy i tekhnologii semanticheskogo veba, Vol. 1. S. 109-116.(In Russian) = Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора // инженерия знаний и технологии семантического веба. Т. 1. 2012. С. 109-116.