УДК 004.912 DOI: 10.18698/2541-8009-2022-05-794

СОЗДАНИЕ СИСТЕМЫ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ОТЗЫВОВ НА РУССКОМ ЯЗЫКЕ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Т.А. Кузнецов tima.kuznetsov1507@gmail.com
С.И. Гавриленков gavrilenkovsergei@bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федераци

Аннотация

В современной высококонкурентной среде предприятия могут повысить свою гибкость и рентабельность благодаря проведению аналитических исследований текстовых отзывов потребителей. В рамках этих исследований одной из первоначальных задач является определение класса тональности текстового отзыва для понимания общей оценки продукта потребителем. В статье рассмотрена задача классификации текстовых отзывов по классам сентиментов с применением методов машинного обучения. В ходе решения изучены и применены методы векторизации текстовых данных. Проведен сравнительный анализ алгоритмов классификации по классам тональности: алгоритма случайного леса, метода опорных векторов, наивного байесовского классификатора. Выбран алгоритм, показывающий наилучшие показатели по метрикам оценки качества модели классификации. Получена подсистема классификации текстовых отзывов, автоматизирующая процесс анализа текстовых данных в рамках исследования продукции, производимой предприятием.

Ключевые слова

Машинное обучение, Индустрия 4.0, обработка естественного языка, классификация текстовых отзывов, автоматизация обработки текстовых отзывов, анализ тональности текстовых отзывов, векторизация текстовых данных, наивный байесовский классификатор

Поступила в редакцию 19.04.2022

Введение. В современном мире особую актуальность получила концепция «Индустрия 4.0», в рамках которой деятельность предприятий трансформируется в результате автоматизации технологических и административных процессов, интеграции информационных технологий на уровне принятия решений. Производство в рамках «Индустрии 4.0» предполагает гиперперсонализированность, что означает исследование рынка реализации продукции, потребностей конечных пользователей не только на стадии проектирования продукта, но и после выпуска. Таким образом можно своевременно вносить изменения в создаваемый предприятием продукт.

Одним из источников получения информации о продукте служат отзывы потребителей. Проведение аналитического исследования отзывов покупателей

о товаре является важнейшей задачей отдела продуктовой аналитики современного предприятия. Многие предприятия реализуют анализ текстовых отзывов в ручном режиме, что приводит к увеличению времени решения задачи и вероятности возникновения ошибки, поскольку человек не способен обрабатывать большое количество информации с постоянным уровнем эффективности. В подобных исследованиях первоначально необходимо классифицировать входные текстовые отзывы по классам тональности: положительный, отрицательный, нейтральный класс. В настоящее время эту задачу можно решать автоматически с применением методов машинного обучения. Данная статья посвящена автоматизации процесса анализа отзывов с применением технологий машинного обучения. В рамках исследования была решена задача автоматической классификации текстовых отзывов по классам тональности. Эта задача относится к классу задач обработки естественного языка, или NLP-задач (Natural Language Preprocessing).

Литературный обзор. Исследователи из разных стран по-разному решают задачи обработки естественного языка. Б. Лю с соавторами в [1] рассматривали области применения обработки естественного языка в электронной коммерции и других областях бизнеса. В статье [2] Б. Паг с соавторами проводили сравнительный анализ определения тональности отзывов о фильмах человеком и моделью машинного обучения. В эксперименте исходные тексты не обрабатывались. Модель была основана на применении подхода TF-IDF (от англ. Term Frequency — Inverse Document Frequency — частота слова – обратная частота документа), а также наивного байесовского классификатора (Naïve Bayes Classifier) и метода опорных векторов (Support Vector Machine). Как показал эксперимент, метод опорных векторов наилучшим образом подходит для классификации, показав 83%-ную точность (по метрике accuracy[1]) определения класса на валидационной выборке. В [3] Б. Трстеняк с соавторами придерживались схожего порядка обработки, однако для классификации использовали только метод *k* ближайших соседей.

Работа [4] посвящена важности предобработки текста перед построением векторной матрицы и классификацией. Первоначально из текста были удалены HTML-теги, затем удалены служебные части речи и каждое слово было приведено к начальной форме. Затем были использованы частотные модели векторизации текста. Авторы применяли только метод опорных векторов в качестве алгоритма классификации. В результате наибольшую точность показала модель TF-IDF с результатом 78%-ной точности (accuracy). Предобработка текста в [5]

[1] Метрика accuracy (от англ. *accuracy* — точность) показывает отношение количества объектов с верно определенным классом к общему количеству объектов в процессе классификации.

позволила А. Трифани с соавторами добиться 94%-ной точности (accuracy) при использовании метода опорных векторов и 89,5%-ной при использовании наивного байесовского классификатора. В [6] А. Сруджан с соавторами классифицировал отзывы о книгах, полученные от платформы Amazon. В работе текст предварительно обрабатывался, далее была применена модель TF-IDF. Автор сравнивал работу пяти различных алгоритмов классификации: алгоритм *k* ближайших соседей (K-Nearest Neighbors), алгоритм деревьев решений (Decision Trees Classifier), наивный байесовский классификатор, метод опорных векторов, алгоритм случайного леса (Random Forest Classifier). Наилучший результат был получен с помощью алгоритма случайного леса.

Т. Хаки с коллегами в статье [7] утверждают, что полученная ими модель классификации отзывов о продукции сайта Amazon отличается серьезным улучшением качества (до 94 % по метрике accuracy) благодаря предобработке текста и применению метода χ-квадрат. С. Дей с соавторами в статье [8] выполнили сравнительный анализ метода опорных векторов и алгоритма наивного байесовского классификатора при решении задачи классификации отзывов в наборе данных Amazon (существует в открытом доступе). В результате исследования была выявлена более высокая точность с применением алгоритма опорных векторов. Поляков Е. с соавторами в статье [9] обобщили существующие подходы к решению задачу идентификации настроения коротких текстовых сообщений в NLP-задачах. В результате исследования выявлена лучшая по качеству комбинация методов классического машинного обучения: логистическая регрессия, построенная на основе частотной модели с применением лемматизации[1] и стемминга[2]. Показатель AUC ROC[3][4] : 0,87.

А. Двойникова и А. Карпов в статье [10] представили подход к анализу тональности русскоязычных текстовых данных. Авторы отмечают более низкую точность в задачах классификации для русскоязычных текстов, чем для англоязычных, связывая это со сложностью языка. С. Сметанин с соавторами в [11] анализирует задачу классификации отзывов для русского языка комплексно: перечисляет существующие подходы к решению и указывает на трудности, возникающие в процессе анализа текстовых данных на русском языке. Е. Котель-

---

[1] Лемматизация (от англ. *lemmatization*) — процесс приведения словоформы к нормальной форме.

[2] Стемминг (от англ. *stemming*) — процесс нахождения основы слова.

[3] Кривая рабочей характеристики приемника (ROC-кривая, от англ. *Receiver Operating Characteristic*) — график, позволяющий оценить качество бинарной классификации. Показывает отношение в классификации между объектами, отнесенными к положительному классу.

[4] AUC ROC (от англ. *Area Under ROC curve*) — показатель, равный площади по ROC кривой.

ников с соавторами в [12] проводят сравнительный анализ наборов обучающих выборок на русском языке для классификации текстовых данных по тональности. Исследование показало, насколько качество обучающей выборки влияет на решение задачи классификации. Этот факт подтверждается исследованием А. Хассана и соавторов [13], в рамках которого был проведен масштабный анализ задачи распознавания настроения текста. Авторы отмечали, что для многих языков до сих пор нет достаточного количества качественно размеченных обучающих данных.

В. Рыбаков и А. Малафеев в [14] проводят определение тональности отзывов об отелях на русском языке. В результате классификации точность для каждой из группы отзывов (о номере, о локации, о сервисе) была в среднем 69 %. А. Звонарев с соавторами в [15] проводят сравнительный анализ логистической регрессии, XGBoost и классификации на основе нейронных сетей для текстовых данных на русском языке. Авторы столкнулись с трудностью определения гиперпараметров для XGBoost а также установили, что логистическая регрессия показывает лучшее качество при малом времени обучения.

Как видно из обзора литературы, посвященной теме классификации тональности текстовых данных, проведено множество исследований. В представленной работе выполнена классификация тональности отзывов о продукции на русском языке. Эта задача на данный момент уникальна, ее решение внесет вклад в методику автоматической обработки текстовых отзывов на русском языке.

Таким образом, целью данного исследования было создание автоматизированной системы классификации отзывов о продукте предприятия для русскоязычных отзывов. Рассмотрим архитектуру системы и укрупненный алгоритм работы более подробно. Далее представим лингвистические модели, с помощью которых решалась поставленная задача. Затем определим процедуры предобработки текстовых данных для последующей классификации и приведем данные об обучающей и валидационных выборках и алгоритмах классификации, используемых в исследовании. По итогам исследования проанализируем полученные результаты и сформулируем направления дальнейшей работы.

Методология. ***Системная архитектура и алгоритм работы программы в целом.*** Архитектура системы представлена на рис. 1. Концептуально система состоит из модуля для сбора начальных данных с помощью web-парсера[1] либо

[1] Синтаксический анализ (жарг. *парсинг* от англ. *parsing*) — процесс сопоставления линейной последовательности лексем естественного или формального языка с его формальной грамматикой. Применяется для автоматического сбора информации в сети Интернет.

CRM[1]-системы. Собранные начальные данные проходят предобработку в соответствующем модуле, после чего модуль машинного обучения осуществляет классификацию отзывов по тональности. Результаты классификации представляются продуктовому аналитику в виде инструментальной панели — дашборда[2].

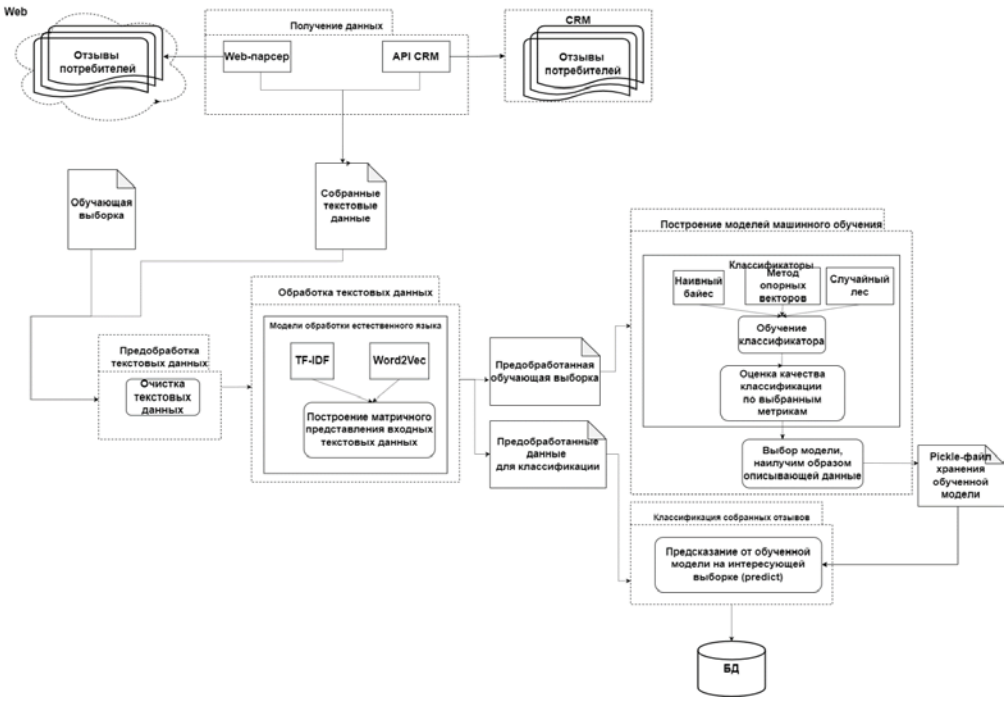


Рис. 1. Архитектуры автоматизированной системы анализа отзывов:
БД — база данных предприятия; API CRM — программный интерфейс системы управления взаимоотношением с клиентом

Для сбора данных был создан web-парсер, а также выполнено подключение к CRM-системе предприятия посредством API[3]-интерфейса. Таким образом была собрана тестовая выборка, для которой необходимо выполнить классификацию.

Для обучения и валидации модели классификации использовали стороннюю обучающую выборку. Последовательность предобработки данных на обучаю-

[1] Система управления взаимоотношениями с клиентом (CRM, от англ. *Customer Relationship Management*) — прикладное программное обеспечение для автоматизированного взаимодействия с заказчиком (клиентом).

[2] От англ. *dashboard* — приборная панель.

[3] Программный интерфейс приложения (API от англ. *application program interface*) — описание процедур, посредством которых осуществляется программное взаимодействие.

щей, валидационной и тестовых выборках была сохранена. После векторизации было выполнено сравнение трех классификаторов для выбора наилучшего решения в рамках задачи.

Представление текстовых данных. Лингвистические модели. В данном исследовании для преобразования корпуса текстовых документов в числовую матрицу были использованы две лингвистические модели: TF-IDF и Word2Vec. Модель TF-IDF — это вариация частотной модели BoW[1], для которой создается словарь уникальных слов корпуса. Числовая матрица, соответствующая корпусу документов, заполняется значением TF-IDF для каждого слова документа в соответствии с вхождением слов в документ в соответствии с формулой

$$TF\text{-}IDF = TF(t,d) \times IDF(t);$$

$$TF(t,d) = \frac{n_t}{\sum_k n_k};$$

$$IDF(t,d) = \lg \frac{|D|}{\left|\left\{d_i \in D \mid t \in d_i\right\}\right|},$$

где n_t — число вхождений слова t в документ; $\sum_k n_k$ — число слов в данном документе; $|D|$ — число документов в коллекции; $\left|\left\{d_i \in D | t \in d_i\right\}\right|$ — число документов из коллекции D, в которых встречается документ t (когда $n_t \neq 0$).

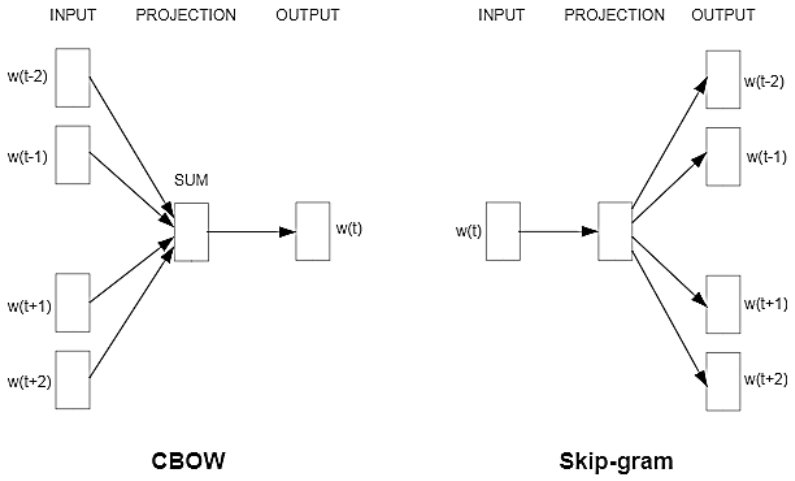


Рис. 2. Алгоритм Word2Vec:
Input — входные параметры; Projection — проекция; Output — выходные параметры

[1] От англ. *bag of words* — мешок слов.

Источник: [16]

Модель Word2Vec — это языковая модель дистрибутивной семантики, основанная на работе нейронных сетей. По данному слову документа нейронная сеть предсказывает окружение этого слова. В результате для каждого документа можно получить многомерный числовой вектор, в котором каждое слово есть вектор вероятности. Принцип работы метода показан на рис. 2. Рассмотрим его более подробно. Задача алгоритма CBOW[1] заключается в том, чтобы предсказать слово $w(t)$, имея в качестве входных параметров контекстные слова ($w(t-1)$, $w(t-2)$, …, $w(t-n)$). Алгоритм Skip-gram[2] предсказывает контекстные слова, имея данное слово в качестве входного параметра.

Предобработка текстовых данных. Для обучения модели классификации необходима размеченная выборка текстовых отзывов. Для исследования использовали обучающую выборку RuReviews, которая находится в открытом доступе на портале github.com проекта Natasha [17]. Обучающая выборка состояла из 90 000 отзывов, распределенных по трем классам: положительный, нейтральный, отрицательный. Выборка была сбалансирована. Было применено предварительное разбиение выборки на обучающую и валидационную в соотношении 75 : 25. Тестовая выборка из 18 000 текстовых отзывов, для которых обученный классификатор определял тональность текстового, была предоставлена коммерческим предприятием и не может быть опубликована в статье в связи с политикой конфиденциальности правообладателя.

Для уменьшения размерности матрицы векторов, представляющей данные для классификации, необходимо предобрабатывать входные текстовые данные. Предобработка текстовых данных состояла из следующих пунктов:

1) приведение слов к нижнему регистру;

2) удаление знаков пунктуации, пробелов, цифр, HTML-тегов, специальных символов (например, скобок [],{},());

3) удаление стоп-слов (т. е. слов, не несущих смысловой ценности для классификации: междометия, служебные части речи);

4) стемминг — выделение корня слова.

Алгоритмы классификации. В результате анализа литературы, посвященной задаче классификации тональности текстовых отзывов, был выделен круг классификаторов для сравнения: алгоритм случайного леса, метод опорных векторов, наивный байесовский классификатор. Алгоритм случайного леса — это ансамбль деревьев решений, который использует предсказания большого количества деревьев для формирования финального ответа. Метод наивного байесовского классификатора — это вероятностный метод классификации, который

[1] От англ. *continuous bag of words* — непрерывный мешок слов.

[2] От англ. *skip-gram* — словосочетание с пропуском.

часто используется в задачах классификации текстовых данных благодаря небольшому необходимому объему обучающей выборки. Метод опорных векторов является одним из самых популярных методов классификации в текстовых задачах. В данном алгоритме определяется гиперплоскость, наилучшим образом разделяющая точки путем вычисления расстояний от точек до предполагаемой плоскости.

Результаты и обсуждение. В эксперименте решали задачу классификации по трем классам тональности отзывов: положительный, отрицательный, нейтральный. Как было упомянуто выше, процедуры обработки текстовых данных для обучающей и тестовой выборок были одинаковыми. Для представления текстовых данных использовали подсчет TF-IDF-значений и алгоритм Word2Vec. Затем проводили обучение трех упомянутых выше классификаторов для каждого способа векторизации и выполняли сравнительный анализ.

Метриками качества для классификаторов были выбраны: F1-мера[1], точность (accuracy), ROC-кривые.

Полученные значения F1 для каждого класса представлены в табл. 1. Общие значения точности (accuracy) приведены в табл. 2.

Таблица 1

Значения F1 для каждого класса

Класс	TF-IDF			Word2Vec		
	Наивный байесовский классификатор	Метод опорных векторов	Алгоритм случайного леса	Наивный байесовский классификатор	Метод опорных векторов	Алгоритм случайного леса
Положительный	0,84	0,84	0,80	0,69	0,77	0,78
Отрицательный	0,68	0,70	0,69	0,59	0,65	0,65
Нейтральный	0,64	0,60	0,60	0,53	0,57	0,58

[1] F1-мера (от англ. *F1-score*) — среднее гармоническое метрик полноты (от англ. *recall*) и точности (от англ. *precision*).

Полнота (англ. *recall*) — метрика, показывающая долю верно определенных объектов положительного класса среди всех объектов положительного класса (могут быть отнесены как к положительному так и ошибочно к отрицательному).

Точность (англ. *precision*) — метрика, показывающая долю верно определенных объектов положительного класса среди всех объектов (положительного и отрицательного классов), отнесенных к положительному классу.

В общем	0,72	0,71	0,69	0,61	0,66	0,67

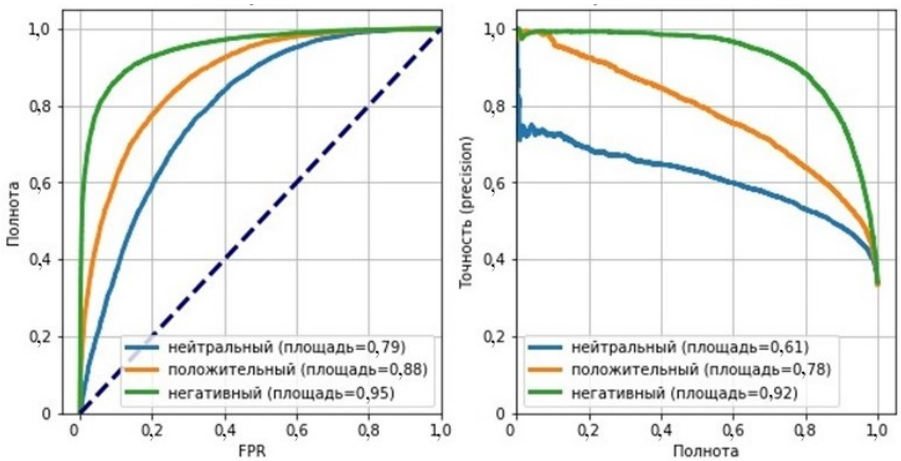
Как видно из табл. 1, для метода TF-IDF наилучшим образом разделяет классы наивный байесовский классификатор. При использовании алгоритма Word2Vec предпочтительным будет использование алгоритма случайного леса. Это может быть связано с тем, что матрица TF-IDF достаточно разреженная.

Таблица 2

Точность (accuracy) в общем по классам, %

Классификатор	TF-IDF	Word2vec
Наивный байесовский классификатор	71,9	60,1
Метод опорных векторов	71,3	66,3
Алгоритм случайного леса	69,7	66,8

Согласно табл. 2 в целом задачу классификации отзывов по трем классам тональности с наибольшим значением точности позволяет решать наивный байесовский классификатор. Метод Word2Vec в данном случае показывает меньшую точность, поскольку для использования его результатов в задаче классификации необходимо выполнять трансформацию многомерного вектора в одномерный, часть семантической информации при этом теряется. Метод Word2Vec предпочтительнее использовать в глубоком обучении с применением нейронных сетей.



а *б*

Рис. 3. ROC Графические метрики оценки качества классификации для TF-IDF и наивного байесовского классификатора:
а — ROC-кривые; *б* — кривые точности-полноты

ROC-кривые (ROC-кривые строятся в координатах полнота -FPR[1]) и кривые точности-полноты (кривые точности полноты строятся в координатах точность-полнота) для метода опорных векторов при векторизации TF-IDF представлены на рис. 3, *а*, *б*. Графиках отображены для каждого класса (классы выделены цветом). На рис. 3 а отображена прямая (пунктирная линяя), показывающая поведение классификатора при случайной оценке класса. На рис. 3а и 3б указаны значения площади под соответствующими кривыми.

Как можно судить по рис. 3, с наибольшей уверенностью классификатор определяет класс положительных отзывов. Определение класса нейтральных отзывов остается трудной задачей, разница в уверенности по сравнению с положительным классом ощутима.

Заключение. Анализ текстовых отзывов является важной задачей, определяющей успешное функционирование предприятия в условиях развития электронной коммерции и трансформации деятельности в рамках концепции «Индустрия 4.0». Результаты анализа отзывов позволяют не только определить положительные и отрицательные стороны производимой продукции, но и внести вклад в разрабатываемые рекомендательные системы для покупателей, и служат источником информации о пожеланиях потребителя. В данной статье выполнен обзор разработанной модели многоклассовой классификации для текстовых отзывов на русском языке о продукции предприятия, реализуемой предприятием на российском рынке. В исследовании был проведен сравнительный анализ двух основных подходов векторизации текстовых данных, а также работы трех классификаторов. В результате исследования можно установить, что наивный байесовский классификатор в сочетании с TF-IDF-методом показывают на приведенной валидационной выборке наибольшее качество классификации (по метрике точности(accuracy)). Для определения слабых мест классификации приведены результаты классификации по каждому классу (по метрике F1-мера). На основе приведенных результатов можно установить, что задача определения нейтрального класса отзывов является наиболее трудной в данном случае. Это связано с особенностями русского языка, а также со спецификой нейтральных отзывов потребителей, содержащих слова-признаки как положительного, так и отрицательного классов отзывов.

Возможным решением проблемы определения нейтральных отзывов (и, как следствие, способом повышения качества классификации модели) может быть применение методов глубокого обучения. В частности, рассмотренный в статье метод векторизации Word2vec можно применять в сочетании с рекуррентной

[1] FPR (от англ. *false positive rate*) — количественная характеристика классификатора, показывающая отношения количества неверно определенных объектов положительного класса ко всем объектам отрицательного класса.

нейронной сетью. В дальнейших исследованиях акцент будет сделан на улучшении качества классификации.

Литература

[1] Liu B., Zhang L. A survey of opinion mining and sentiment analysis. In: *Mining text data*. Springer, 2021, pp. 415–463. DOI: https://doi.org/10.1007/978-1-4614-3223-4_13

[2] Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. *Proc. EMNLP,* 2002, pp. 79–86. DOI: https://doi.org/10.3115/1118693.1118704

[3] Trstenjak B., Mikac S., Donko D. KNN with TF-IDF based framework for text categorization. *Procedia Eng.*, 2014, vol. 69, pp. 1356–1364. DOI: https://doi.org/10.1016/j.proeng.2014.03.129

[4] Haddi E., Liu X., Shi Y. The role of text pre-processing in sentiment analysis. *Procedia Comput. Sci.*, 2013, vol. 17, pp. 26–32. DOI: https://doi.org/10.1016/j.procs.2013.05.005

[5] Tripathy A., Agrawal A., Rath S.K. Classification of sentimental reviews using machine learning techniques. *Procedia Comput. Sci.*, 2015, vol. 57, pp. 821–829. DOI: https://doi.org/10.1016/j.procs.2015.07.523

[6] Srujan K.S., Nikhil S.S., Raghav Rao H. et al. Classification of Amazon book reviews based on sentiment analysis. In: *Information systems design and intelligent applications*. Springer, 2018, pp. 401–411. DOI: https://doi.org/10.1007/978-981-10-7512-4_40

[7] Haque T.U., Saber N.N., Shah F.M. Sentiment analysis on large scale Amazon product reviews. *Proc. ICIRD*, 2018. DOI: https://doi.org/10.1109/ICIRD.2018.8376299

[8] Dey S., Wasif S., Tonmoy D.S. et al. A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. *Proc. IC3A*, 2020, pp. 217–220. DOI: https://doi.org/10.1109/IC3A48958.2020.233300

[9] Поляков Е.В., Восков Л.С., Абрамов П.С. и др. Исследование обобщенного подхода к решению задач анализа настроений коротких текстовых сообщений в задачах обработки естественного языка. *Информационно-управляющие системы*, 2020, № 1, с. 2–14. DOI: https://doi.org/10.31799/1684-8853-2020-1-2-14

[10] Двойникова А.А., Карпов А.А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных. *Информационно-управляющие системы*, 2020, № 4, с. 20–30. DOI: https://doi.org/10.31799/1684-8853-2020-4-20-30

[11] Smetanin S. The applications of sentiment analysis for Russian language texts: current challenges and future perspectives. *IEEE Access*, 2020, vol. 8, pp. 110693–110719. DOI: https://doi.org/10.1109/ACCESS.2020.3002215

[12] Kotelnikov E., Peskisheva T., Kotelnikova A. et al. A comparative study of publicly available Russian sentiment lexicons. *Proc. AINL 2018*. Springer, 2018, pp. 139–151. DOI: https://doi.org/10.1007/978-3-030-01204-5_14

[13] Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng. J.*, 2014, vol. 5, no. 4, pp. 1093–1113. DOI: https://doi.org/10.1016/j.asej.2014.04.011

[14] Rybakov V., Malafeev A. Aspect-based sentiment analysis of Russian hotel reviews. *Proc. AIST-SUP*, 2018, pp. 75–84.

[15] Zvonarev A., Bilyi A. A comparison of machine learning methods of sentiment analysis based on Russian language twitter data. *Proc. MICSECS*, 2019.
URL: https://dblp.org/rec/conf/micsecs/ZvonarevB19.html (дата обращения: 15.05.2022).
[16] Mikolov T., Chen K., Corrado G. et al. Efficient estimation of word representations in vector space. *Proc. Workshop at ICLR,* 2013.
DOI: https://doi.org/10.48550/arXiv.1301.3781
[17] Natasha: tools for Russian NLP. *github.com: веб-сайт*. URL: https://github.com/natasha (дата обращения: 15.05.2022).

Кузнецов Тимофей Александрович — студент кафедры «Компьютерные системы автоматизации производства», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Научный руководитель — Гавриленков Сергей Игоревич, ассистент кафедры «Компьютерные системы автоматизации производства», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ссылку на эту статью просим оформлять следующим образом:
Кузнецов Т.А. Создание системы автоматической классификации текстовых отзывов на русском языке с помощью машинного обучения. *Политехнический молодежный журнал*, 2022, № 05(70). http://dx.doi.org/10.18698/2541-8009-2022-05-794

DEVELOPMENT OF AUTOMATIC CLASSIFICATION SYSTEM FOR TEXTUAL FEEDBACK IN RUSSIAN LANGUAGE USING MACHINE LEARNING

T.A. Kuznetsov tima.kuznetsov1507@gmail.com
S.I. Gavrilenkov gavrilenkovsergei@bmstu.ru

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

In today's highly competitive environment enterprises could increase their flexibility and profitability by conducting analytical studies of textual customer feedback. One of the primary objectives of these studies is to determine the sentiment class of textual feedback in order to understand the consumer's overall evaluation of the product. This paper considers the task of classifying textual feedback into sentiment classes using machine learning techniques. Text data vectorization methods are studied and applied while solving the task. A comparative analysis of algorithms for classifying by sentiment classes is carried out: random forest algorithm, support vector machine, and naive Bayes classifier. The algorithm showing the best performance on quality evaluation metrics of the classification model is chosen. The subsystem of textual feedback classification that automates the process of text mining in the framework of enterprise product research is obtained.

Keywords

Machine learning, Industry 4.0, natural language processing, textual feedback classification, automation of textual feedback processing, sentiment analysis of textual feedback, text data vectorization, naive Bayes classifier

Received 19.04.2022

References

[1] Liu B., Zhang L. A survey of opinion mining and sentiment analysis. In: *Mining text data*. Springer, 2021, pp. 415–463. DOI: https://doi.org/10.1007/978-1-4614-3223-4_13

[2] Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. *Proc. EMNLP*, 2002, pp. 79–86. DOI: https://doi.org/10.3115/1118693.1118704

[3] Trstenjak B., Mikac S., Donko D. KNN with TF-IDF based framework for text categorization. *Procedia Eng.*, 2014, vol. 69, pp. 1356–1364. DOI: https://doi.org/10.1016/j.proeng.2014.03.129

[4] Haddi E., Liu X., Shi Y. The role of text pre-processing in sentiment analysis. *Procedia Comput. Sci.*, 2013, vol. 17, pp. 26–32. DOI: https://doi.org/10.1016/j.procs.2013.05.005

[5] Tripathy A., Agrawal A., Rath S.K. Classification of sentimental reviews using machine learning techniques. *Procedia Comput. Sci.*, 2015, vol. 57, pp. 821–829. DOI: https://doi.org/10.1016/j.procs.2015.07.523

[6] Srujan K.S., Nikhil S.S., Raghav Rao H. et al. Classification of Amazon book reviews based on sentiment analysis. In: *Information systems design and intelligent applications*. Springer, 2018, pp. 401–411. DOI: https://doi.org/10.1007/978-981-10-7512-4_40

[7] Haque T.U., Saber N.N., Shah F.M. Sentiment analysis on large scale Amazon product reviews. *Proc. ICIRD*, 2018. DOI: https://doi.org/10.1109/ICIRD.2018.8376299

[8] Dey S., Wasif S., Tonmoy D.S. et al. A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. *Proc. IC3A*, 2020, pp. 217–220. DOI: https://doi.org/10.1109/IC3A48958.2020.233300

[9] Polyakov E.V., Voskov L.S., Abramov P.S. et al. Generalized approach to sentiment analysis of short text messages in natural language processing. *Informatsionno-upravlya-}yushchie sistemy* [Information and Control Systems], 2020, no. 1, pp. 2–14. DOI: https://doi.org/10.31799/1684-8853-2020-1-2-14 (in Russ.).

[10] Dvoynikova A.A., Karpov A.A. Analytical review of approaches to Russian text sentiment recognition. *Informatsionno-upravlyayushchie sistemy* [Information and Control Systems], 2020, no. 4, pp. 20–30. DOI: https://doi.org/10.31799/1684-8853-2020-4-20-30 (in Russ.).

[11] Smetanin S. The applications of sentiment analysis for Russian language texts: current challenges and future perspectives. *IEEE Access*, 2020, vol. 8, pp. 110693–110719. DOI: https://doi.org/10.1109/ACCESS.2020.3002215

[12] Kotelnikov E., Peskisheva T., Kotelnikova A. et al. A comparative study of publicly available Russian sentiment lexicons. *Proc. AINL 2018*. Springer, 2018, pp. 139–151. DOI: https://doi.org/10.1007/978-3-030-01204-5_14

[13] Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng. J.*, 2014, vol. 5, no. 4, pp. 1093–1113. DOI: https://doi.org/10.1016/j.asej.2014.04.011

[14] Rybakov V., Malafeev A. Aspect-based sentiment analysis of Russian hotel reviews. *Proc. AIST-SUP*, 2018, pp. 75–84.

[15] Zvonarev A., Bilyi A. A comparison of machine learning methods of sentiment analysis based on Russian language twitter data. *Proc. MICSECS*, 2019. URL: https://dblp.org/rec/conf/micsecs/ZvonarevB19.html (accessed: 15.05.2022).

[16] Mikolov T., Chen K., Corrado G. et al. Efficient estimation of word representations in vector space. *Proc. Workshop at ICLR*, 2013. DOI: https://doi.org/10.48550/arXiv.1301.3781

[17] Natasha: tools for Russian NLP. *github.com: website*. URL: https://github.com/natasha (accessed: 15.05.2022).

Kuznetsov T.A. — Student, Department of Computer Systems of Production Automation, Bauman Moscow State Technical University, Moscow, Russian Federation.

Gavrilenkov S.I. — Assistant, Department of Computer Systems of Production Automation, Bauman Moscow State Technical University, Moscow, Russian Federation.

Please cite this article in English as:
Kuznetsov T.A. Development of automatic classification system for textual feedback in russian language using machine learning. *Politekhnicheskiy molodezhnyy zhurnal* [Politechnical student journal], 2022, no. 05(70). http://dx.doi.org/10.18698/2541-8009-2022-05-794.html (in Russ.).