УДК 004.891.3; 519.254

Л.Ю. Осипова
Байкальский государственный университет,
г. Иркутск, Российская Федерация

В.В. Братищенко
Байкальский государственный университет,
г. Иркутск, Российская Федерация

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ЗАЯВОК СЛУЖБЫ ПОДДЕРЖКИ ПОЛЬЗОВАТЕЛЕЙ

Аннотация. Предлагается использовать автоматическую классификацию обращений в службу поддержки пользователей для выбора подразделения по обработке обращения. Выбраны метод параметрического описания обращения, алгоритмы классификации, программное обеспечение классификации. По выборке обращений выполнена настройка классификаторов, определен классификатор, обеспечивающий лучшее качество классификации, исследована точность классификации в зависимости от максимальной вероятности класса. Предложены рекомендации по применению автоматической классификации обращений.

Ключевые слова. Автоматическая классификация текста, text mining, bag of words, обработка естественного языка.

Информация о статье. Дата поступления: 16 февраля 2021 г.

L.Yu. Osipova
Baikal State University,
Irkutsk, Russian Federation

V.V. Bratishenko
Baikal State University,
Irkutsk, Russian Federation

AUTOMATIC CLASSIFICATION OF USER SUPPORT APPLICATIONS

Abstract. The article describes the model for analyzing the input flow of customer support applications to automatically determine the reasons for those applications. Algorithms for the classification of applications, according to the corpus of references, were selected, the classifiers were tuned, a classifier was determined that would provide the best quality of classification, the classification accuracy was investigated depending on the maximum probability of a class, and recommendations for the use of automatic classification of applications were developed.

Keywords. Automatic classification of applications, text mining, bag of words, natural language processing.

Article info. Received 16 February, 2021.

Автоматизированная система управления единой службой поддержки пользователей (АСУ ЕСПП)[1] информационных технологий обеспечивает взаимодействие со всеми пользователями в рамках многоуровневой поддержки предоставления ИТ услуг в ОАО «Российские железные дороги». Диспетчеры контакт-центра осуществляют регистрацию, классификацию и назначение обращений в соответствующую рабочую группу специалистов для выполнения необходимых работ в рамках предоставления ИТ-услуг. В системе регистрируется около 20 000 обращений пользователей ежедневно, а также выполняется порядка 10 000 плановых и регламентных работ по обслуживанию специалистами ИТ-инфраструктуры. В условиях оперативной работы и большого количества информационных ресурсов у специалиста контакт-центра зачастую нет возможности однозначно определить обращение в соответствующую группу поддержки. Неверная классификация и маршрутизация обращения влечет увеличение времени ее обработки, нарушение соглашения об уровне ИТ-обслуживания, а, следовательно, приводит к потерям для бизнеса.

В работе для решения этой проблемы предлагается использование автоматической классификации обращений с применением технологий интеллектуального анализа данных. Выявление и определение причины на момент поступления заявки в рабочую группу позволит повысить качество и скорость обработки обращений.

Автоматическая классификация текстовых данных является сложной трудоемкой задачей, прежде всего, в связи с необходимостью обработки естественного языка. Обращения в службу поддержки формулируются пользователем в свободной форме, не имеют четкой структуры и формальных признаков идентификации проблемы. Классификация текстов является одной из задач интеллектуальной технологии Text Mining — анализа текстовой информации — алгоритмического процесса обнаружения неизвестных ранее знаний из текста, а также выявления основных понятий и взаимосвязей между ними. Процесс извлечения новых знаний из текстовой информации является нетривиальным и достаточно трудоемким [1].

Процедура построения автоматического классификатора состоит из следующих этапов: предварительная обработка текстов, преобразование текстов, формирование параметрического описания текстов, выбор и применение методов классификации, оценка работы классификатора [2].

Для моделирования был использован инструмент с открытым исходным кодом Orange[2], предоставляющий широкие воз-

[1] Управление обращениями в ОАО «Российские железные дороги».- URL: https://digdes.ru/project/rzhd-upravlenie-obrashhenijami

[2] Orange. Documentation.- URL: https://orangedatamining.com/docs/

можности для анализа данных и визуализации результатов обработки данных. Исходная выборка для настройки классификации включала в себя 521 обращение с правильно определенным классом и следующими атрибутами: подробное описание проблемы, предложенное решение, оказанная услуга, код услуги.

На этапе анализа обращений были выделены следующие классы обращений:

– ведение нормативно-справочной информации;
– методологический вопрос применения ИТ-сервисов;
– ведение учетных записей пользователей;
– технологическая консультация в рамках инструкции;
– технологическая консультация по вопросам, не отраженным в инструкции;
– техническая причина сбоя программно-технического комплекса или сети передачи данных;
– ошибка прикладного программного обеспечения автоматизированной системы;
– установка или настройка автоматизированной системы на рабочее место;
– ошибка внешней или смежной системы;
– ошибка пользователя;
– техническое обслуживание рабочего места.

В ходе предварительной обработки текстовых данных выполнено преобразование исходного текста в форму, удобную для применения алгоритмов Text Mining и включающее следующие этапы:

1) приведение текста к нижнему регистру;

2) токенизация — разбиение текста на отдельные слова-токены;

3) удаление стоп-слов — служебных частей речи и других коротких слов, которые не несут смысловой нагрузки, что позволяет сократить объём текста и увеличить его смысловую значимость;

4) удаление всех нерелевантных символов — знаков пунктуации и цифр;

5) лемматизация текста — приведение слова к его словарной форме, то есть выделение у заданного слова словарной формы — леммы. При этом происходит преобразование в исходном тексте грамматической формы слова (падежи, род, число прилагательных, глагольные виды и времена, залоги причастий и так далее) к словарной форме, что позволяет рассматривать слова с одинаковыми леммами как один и тот же термин [3].

На выходе данного этапа формируется последовательность лемм без лишних компонентов, не влияющих на результат анализа.

На этапе параметрического описания текстов использовался метод Bag-of-Words («мешок слов»). Представление текстов в виде «мешка слов» заключается в определении частоты вхождения

каждого слова в каждый текст. Предложенное описание текстов используется для обучения классификатора. Параметрическое описание множества текстов, которое принято называть корпусом, представляют в виде матрицы, в которой строки соответствуют текстам, а столбцы — словам, включенным в тексты корпуса. Элементами матрицы являются частоты слов. К преимуществам данного подхода относится его универсальность, скорость работы и множество вариантов применение словарей или хеш-функций слов [4].

Параметрическое описание корпуса текстов применяется для настройки классификаторов. Для решения задачи классификации обращений в контакт-центр выбраны простые классификаторы [5]: наивный байесовский алгоритм, дерево решений и логистическая регрессия, которые обладают рядом преимуществ: высоким быстродействием, удобством настройки и легкостью адаптации к новым данным.

Наивный байесовский алгоритм использует частоты слов при условии принадлежности текста известному классу. Условные частоты вычисляются по обучающей выборке в процессе обучения классификатора. Далее эти частоты применяются для вычисления апостериорных вероятностей классов текста. Прогноз класса текста определяется по максимальной апостериорной вероятности. При этом предполагается независимость признаков в совокупности. В процессе обработке текстов по методике «мешка слов» такое предположение можно обосновать игнорированием положения и связи слов в текстах. Алгоритм отличается высокой эффективностью обучения и применения и, в то же время, чувствительностью к нулевым частотам слов.

Дерево решений реализует алгоритм распознавания в виде последовательности проверки частот слов в узлах дерева. С одной стороны, дерево решений эффективно решает задачу классификации, с другой — характеризуется сложностью и неоднозначностью построения дерева в процессе обучения, а также возможностью переобучения (достижение хорошей точности в процессе обучения по обучающей выборке и плохой точности в применении классификатора на тестовой выборке).

Логистическая регрессия обосновывается методом максимального логарифма отношения вероятности $p(x)$ принадлежности классу к вероятности $1 - p(x)$ не принадлежности классу в виде линейной функции

$$ln\left(\frac{p(x)}{1-p(x)}\right) = x_0 + b_1 x_1 + \cdots + b_n x_n$$

от вектора $x = (x_0, \ldots, x_n)$ характеристик. Алгоритм, основанный на применении логистической регрессии, обладает хорошими ха-

рактеристиками обучения и классификации. В данном виде логистическая регрессия определена для бинарной классификации, но может применяться и для множественной классификации в виде последовательности бинарных классификаций.

Более сложные алгоритмы классификации, в том числе использующие ансамбли простых, чувствительны к размерности вектора характеристик классифицируемых объектов. Это особенно справедливо для классификации текстов, представленных частотными характеристиками слов, встречающихся в корпусе текстов.

По имеющемуся корпусу обращений выполнено обучение классификаторов по методике кросс-валидации, когда различные части корпуса последовательно выполняют роль обучающей и тестирующей выборок. В результате получены следующие характеристики алгоритмов классификации (рис. 1):

– AUC (Area under ROC — площадь под ROC кривой) — интегральная характеристика качества обучения, чем ближе AUC к единице, тем выше качество классификатора;

– CA (Classification accuracy) — доля правильно классифицированных обращений;

– Precision — доля правильно классифицированных обращений среди всех обращений, отнесенных классификатором к некоторому классу;

– Recall — доля правильно положительно классифицированных обращений среди всех обращений некоторого класса;

– F1 — взвешенное гармоническое среднее показателей precision и recall;

– Specificity — доля обращений, правильно исключенных классификатором из класса, среди всех обращений других классов.

Относительно невысокие показатели точности CA, Precision, Recall, F1 объясняются ориентацией этих характеристик на бинарную классификацию. В случае классификации обращений количество классов больше двух и показатели точности сравнимы со средним значением частоты наиболее вероятного класса. Как видно из рис. 1, наилучший результат был получен при использовании логистической регрессии.

Прогноз класса в алгоритме классификации определяется по наибольшей вероятности. На рис. 2 представлена зависимость

Model	AUC	CA	F1	Precision	Recall	Specificity
Logistic Regression	0,845	0,478	0,471	0,477	0,478	0,939
Tree	0,701	0,342	0,354	0,413	0,342	0,920
Naive Bayes	0,722	0,148	0,160	0,454	0,148	0,968

Рис. 1. Сравнение характеристик алгоритмов

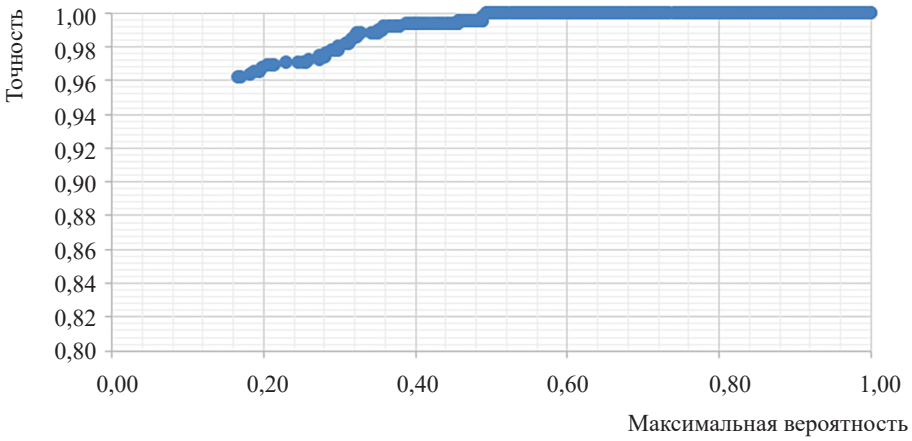

Рис. 2. Точность прогнозируемого класса

точности классификации (доли правильно классифицированных обращений) в зависимости от максимальной вероятности класса. Применение классификатора не гарантирует правильной классификации всех заявок. Однако, при максимальной вероятности класса большей 0,3 количество ошибок будет меньше 2 %.

Описанная технология классификации может быть встроена в ЕСПП в виде правила «автоматическое направление обращения с большой вероятностью класса соответствующим исполнителям и ручная обработка для обращений с недостаточно большой вероятностью класса». Пороговое значение максимальной вероятности — точка отсечения — может быть установлена в процессе эксплуатации автоматической классификации. Целесообразно применять этот алгоритм на этапе создания обращения пользователем до направления обращения в контакт-центр — плохо классифицированные обращения можно автоматически возвращать пользователям для уточнения формулировки. Применение средств автоматической классификации существенно снизит нагрузку на диспетчеров контакт-центра и ускорит обработку обращений.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Катасёв А.С. Симбиоз методов искусственного интеллекта при обнаружении нелинейных зависимостей в базах данных / А.С. Катасёв, Ч.Ф. Ахатова // Информатика, вычислительная техника и инженерное образование. — 2010. — № 2. — С. 46–57.

2. Кириллов А.А. Построение и применение классификатора текстовых обращений в техническую поддержку / А.А. Кириллов, В.И. Виноградов // Моделирование и анализ данных. — 2019. — № 3. — С. 37–42.

3. Bird S. Natural Language Processing with Python / S. Bird, E. Loper, E. Klein. — Cambridge : O'Reilly Media Inc., 2009. — 463 p.

4. Chetviorkin I.I. Sentiment Analysis Track at ROMIP 2011 / I.I. Chetviorkin, P.I. Braslavski, N.V. Loukachevitch // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной Международной конференции «Диалог». — Moscow, 2012. — Вып. 11, т. 2. — С. 1–14.

5. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям : учеб. пособие / Н.Б. Паклин, В.И. Орешков. — 2-е изд., испр. — Санкт-Петербург : Питер, 2013. — 701 с.

REFERENCES

1. Katasev A.S., Akhatova Ch.F. Symbiosis of Artificial Intellect Methods at Nonlinear Dependences Discovering in Databases. *Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie = Computer Science, Computer Science and Engineering Education,* 2010, no. 2, pp. 46–57. (In Russian).

2. Kirillov A.A., Vinogradov V.I. Building and using a classifier of text calls to technical support. *Modelirovanie i analiz dannykh = Modelling and Data Analisis,* 2019, no. 3, pp. 37–42. (In Russian).

3. Bird S., Loper E., Klein E. *Natural Language Processing with Python*. Cambridge, O'Reilly Media Inc., 2009. 463 p.

4. Chetviorkin I.I., Braslavski P.I., Loukachevitch N.V. Sentiment Analysis Track at ROMIP 2011. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialog»]. Moscow, 2012, iss. 11, vol. 2, pp. 1–14.

5. Paklin N.B., Oreshkov V.I. *Biznes-analitika: ot dannykh k znaniyam* [Business Analytics: from the Data to Knowledge]. Saint Petersburg, Piter Publ., 2013. 704 p.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Осипова Людмила Юрьевна — магистрант, кафедра математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, ул. Российская Федерация, e-mail: mobil90@list.ru.

Братищенко Владимир Владимирович — кандидат физико-математических наук, доцент, кафедра математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: vvb@bgu.ru.

INFORMATION ABOUT THE AUTHORS

Lyudmila Yu. Osipova — Graduate Student, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, Russian Federation, e-mail: mobil90@list.ru.

Vladimir V. Bratishenko — Candidate of Physical and Mathematical Sciences, Associate Professor, Department of Mathematical Methods and Digital Technologies, Baikal State University, Russian Federation, e-mail: vvb@bgu.ru.

ДЛЯ ЦИТИРОВАНИЯ

Осипова Л.Ю. Автоматическая классификация заявок службы поддержки пользователей / Л.Ю. Осипова, В.В. Братищенко // System Analysis & Mathematical Modeling. — 2021. — Т. 3, № 1. — С. 45–51.

FOR CITATION

Osipova L.Yu., Bratishenko V.V. Automatic Classification of User Support Applications. *System Analysis & Mathematical Modeling*, 2021, vol. 3, no. 1, pp. 45–51. (In Russian).