УДК 004.912

К.А. Лохачева, Д.И. Парфёнов
Оренбургский государственный университет, г. Оренбург

ИССЛЕДОВАНИЕ КАЧЕСТВА КЛАССИФИКАЦИИ КОРОТКИХ СЛАБОСТРУКТУРИРОВАННЫХ ТЕКСТОВ

В данной статье рассмотрены аспекты классификации коротких слабоструктурированных текстов. Приведены общая постановка задачи, методы векторизации данных и классификации объектов. Описана структура тестовых данных с подробными пояснениями. Сделаны выводы о качестве классификации.

Ключевые слова: классификация текстов; обращения пользователей, LR, SVM, RNN, LSTM.

K.A. Lokhacheva, D.I. Parfenov
Orenburg state university, Orenburg

RESEARCH ON THE QUALITY OF SHORT WEAKLY STRUCTURED TEXTS CLASSIFICATION

In this article the aspects of engine test automation are considered (текст аннотации на английском языке). In this article the aspects of short weakly structured texts classification are being discussed. The problem statement, methods of data vectorization and object classification are given. The structure of the test data is described with detailed explanations. Conclusions about the quality of classification are made.

Keywords: texts classification; issues, LR, SVM, RNN, LSTM.

Введение

Большинство обращений поступают в Service Desk системы в неструктурированном виде, содержат синтаксические и грамматические ошибки, общую лексику, а также не имеют специальных маркеров тематики обращения, либо маркеры носят обобщенный характер. В связи с этим, задача семантического анализа и автоматической классификации обращений пользователей в Service Desk системах представляется актуальной.

Общая постановки задачи звучит следующим образом. Необходимо спроектировать информационную систему для семантического анализа и классификации обращений пользователей в системе Service Desk. Ожидается, что каждая оставленная в системе Service Desk заявка будет проходить предварительную обработку, прежде чем попадет в список заявок, подлежащих исполнению. При этом сам процесс предварительной обработки будет заключаться в семантическом анализе слабо структурированной информации, полученной из заявки, и классификации заявки по типу запроса. После преобразования, заявка

добавляется в список заявок, подлежащих исполнению. Сотрудники могут назначить любую заявку на себя.

Ставится задача выбора оптимального с точки зрения F-меры классификатора.

Основная часть

Математическая постановка задачи выглядит следующим образом. Пусть имеется множество обращений пользователей $R = \{r_1, ..., r_n\}$ и конечное множество заранее определенных классов типов обращений $Tp = \{tp_1, ..., tp_k\}$ Тогда целевая функция $\varphi : R \times Tp \to \{0,1\}$, определяющая, принадлежит ли текстовый документ (то есть обращение пользователя) к данному классу или нет, задается следующим отношениеми:

$$\varphi(r_j, tp_i) = \begin{cases} 0, \text{ если } r_j \notin tp_i \\ 1, \text{ если } r_j \in tp_i \end{cases} \quad (1)$$

Требуется построить классификатор φ', максимально близкиq к φ.

При этом, обращения пользователей будут иметь векторное представление

$$r_j = \left(w_{1j}, ..., w_{mj} \right), \quad (2)$$

где r_j — векторное представление *j*-го обращения ($j = \overline{1,n}$), w_{ij} — вес *i*-го терма в *j*-м обращении, *m* — общее количество различных термов во всех обращениях пользователей.

Под термами будем понимать слова, из которых состоят тексты обращений пользователей.

На сегодняшний день существует несколько типов классифицирующих алгоритмов [1]:

1) вероятностные (Наивный Байесовский метод, NB);

2) метрические (*k*-ближайших соседей, KNN);

3) логические (Деревья решений, DT);

4) линейные (Логистическая регрессия – LR, метод опорных векторов – SVM);

5) методы, основанные на технологии нейронных сетей (рекуррентные нейронные сети – RNN, сверточные нейронные сети – CNN).

У каждой категории методов есть свои плюсы и минусы, подробно описанные в ряде научных работ [1-9].

В рамках нашей модели предполагается исследовать результаты работы следующих моделей: LR, SVM, RNN типа LSTM.

Кроме того, в нашей модели мы рассмотрим 2 метода векторизации пространства:

1) Частота вхождения слова в документе (TF — *term frequency*). Частота вычисляется как отношение числа вхождения слова к общему количеству слов текста:

$$TF(t,r) = \frac{n_t}{\sum_p n_p} \quad (3)$$

где n_t – число вхождений терма *t* в обращении, $\sum_p n_p$ – общее число слов в данном обращении.

Как уже отмечалось ранее, недостатком является несоразмерность оценки для текстов разной длины: недооцениваются длинные документы, так как в них больше слов и средняя частота слов в тексте ниже.

2) TF-IDF – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции:

$$TF - IDF = TF(t,r) \times TF(t,n) \quad (4)$$

где $IDF(t,n)$:

$$IDF(t,n) = \frac{n}{\sum_i r_i} \quad (5)$$

где *n* – число обращений пользователей, $\sum_i r_i$ – число обращений из коллекции, в которых встречается терм *t* ($r_i \in R \mid t \in r_i$).

При этом следует учесть, что LSTM сети принимают на вход данные в TF-векторизации.

F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю

$$F = 2\frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Для проведения эксперимента нами была выбрана среда разработки Python и библиотеки Pandas, Keras, реализующие заявленные методы классификации, ScikitLearn, реализующую заявленные методы векторизации, а также библиотеку NLTK для предобработки данных.

Данные обучающей выборки имеют следующую структуру:

– Issue Type (Тип Запроса) – целевой атрибут;

– Issue key (Ключ) – УИ;

– Summary (Краткое описание) – заголовок запроса;

– Assignee (Исполнитель) – сотрудник системы Service Desk, работающий с заявкой;

– Reporter (Автор) – автор заявки в системе Service Desk;

– Priority (Приоритет) – ранг приоритета срочности заявки;

– Created (Создано) – дата создания заявки;

– Description (Описание) – тело заявки.

При этом для решения задачи классификации обращений по типу запроса нам необходимы предварительно объединенные данные полей Summary и Description, а также метки классов – Issue Type. Обучающая и тестовая выборки составляют, соответственно, 80- и 20-%.

В таблице 1 представлена информация о результатах моделирования.

Таблица 1 – Результаты моделирования

Комбинация алгоритмов	F-мера
TF + LR	0.69
TF-IDF + LR	0.70
TF + SVM	0.68
TF-IDF + SVM	0.70
TF + LSTM	0.38
TF + 2 уровня LSTM	0.324

Исходя из данных таблицы 1 можно заключить, что более простые методы классификации работают лучше, чем более сложные на слабоструктурированных коротких текстах. Кроме того, использование метрики TF-IDF способно хоть и незначительно, но повышать качество классификации.

Список литературы

1. Батура Т.В. Методы автоматической классификации текстов / Т.В. Батура // Программные продукты и системы. – 2017. – Т. 30, № 1. – С. 85–99.

2. Епрев А.С. Автоматическая классификация текстовых документов / А.С. Епрев // Математические структуры и моделирование. – 2010. – №21. – С. 65-81.

3. Мбайкоджи Э. Метод автоматической классификации коротких текстовых сообщений / Э. Мбайкоджи, А.А. Драль, И.В. Соченков // Информационные технологии и вычислительные системы. – 2012. – №3. – С. 93-102.

4. Осипова Ю.А. Применение кластерного анализа методом k-средних для классификации текстов научной направленности / Ю.А. Осипова, Д.Н. Лавров // Математические структуры и моделирование. –2017. – №3(43). – С. 108-121.

5. Отраднов К.К. Модель кластеризации слабоструктурированных текстовых данных / К.К. Отраднов, Д.О. Жуков, О.А. Новикова // Современные информационные технологии и ИТ-образование. Том 13, № 3. –2017. – С.100-113.

6. Шаграев А.Г. Трансдуктивное обучение логистической регрессии в задаче классификации текстов / А.Г. Шаграев, И.А. Бочаров, В.Н. Фальк // Программные продукты и системы. – 2014. – №2. – С. 114 – 118.

7. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: a survey / W. Medhat, A. Hassan, H. Korashy // Ain Shams Eng. Jour. – 2014. – №. 5. – С. 1093–1113.

8. Tarasov D.S. Deep recurrent neural networks for multiple language aspect-based sen-timent analysis / D.S. Tarasov // Computational Linguistics and Intellectual

Technologies. In Proc. Annual Intern. Conf. «Dialogue-2015».– 2015. – №. 2, 14 (21). – С. 65–74.

9. Xiang Zhang, Junbo Zhao, Yann LeCun Character-level convolutional networks for text classification // Proc. Neural Inform. Processing Systems Conf. (NIPS 2015). – Montreal, Canada. – 2015.

Сведения об авторах

Парфёнов Денис Игоревич – кандидат технических наук, доцент кафедры прикладной математики, Оренбургский государственный университет, Оренбург, email: parfenovdi@mail.ru

Лохачева Ксения Алексеевна – магистрант Оренбургского государственного университета, Оренбург, email: ksenia.lohacheva.97@mail.ru

About the authors

Parfenov Denis Igorevich - Candidate of Engineering Sciences, associate professor of the Applied Mathematics department, Orenburg state university, Orenburg, email: parfenovdi@mail.ru

Lokhacheva Ksenia Alekseevna - Student of Orenburg state university, Orenburg, email: ksenia.lohacheva.97@mail.ru