УДК 004.421

АЛГОРИТМ КЛАССИФИКАЦИИ ТЕКСТОВ

И. Ю. Татуревич, О. А. Кравченко

Учреждение образования «Гомельский государственный технический университет имени П. О. Сухого», Республика Беларусь

Классификация текста необходима для решения таких задач, как борьба со спамом, распознавание эмоциональной окраски текстов, разделение сайтов по тематическим разделам, персонификация рекламы. Классификация текстов является задачей семантического анализа текстов, имеющей обширную область применения: психология, политология, филология, торговля, проектирование поисковых и SEO-алгоритмов, искусственного интеллекта и текстовых фильтров [1].

В качестве задачи классификации текста рассматривается задача по фильтрации спама в базе данных смс-сообщений [2]. Спам (англ. spam) – массовая рассылка корреспонденции рекламного характера лицам, не выражавшим желания ее получать. Доля спама в мировом почтовом траффике составляет от 60 до 80 %. Это говорит о том, что исследования новых алгоритмов для выявления такого рода сообщений являются актуальными. С учетом роста влияния информационных технологий на многие аспекты жизни человека также необходимо решать такие задачи с точки зрения скорости обработки таких сообщений, точности их определения и возможности интеграции данного функционала с другими программами.

Разработанный алгоритм работы классификатора состоит в следующем. Исходный массив данных разбивается на слова или предложения (токенизируется). Составляется таблица, в которую включаются сообщения, сведения о принадлежности к классу «спам» или «не-спам», количество разных слов сообщения. Затем происходит преобразование полученных данных в числовую форму с использованием модели *bag-of-words*. Следующий шаг алгоритма – обучение функции классификации тренировочным набором данных: из сообщений выделяются слова, которые являются ключевыми характеристиками для классификации сообщений. При появлении в системе нового сообщения из него также будут выделены слова. При этом, если данные слова не встретятся в составленной ранее таблице, они будут исключены из классификации и будет подсчитано общее количество оставшихся слов. Вероятность того, что сообщение будет отнесено к одному из классов, равна произведению вероятностей вхождения найденных слов в этот класс. Необходимо рассчитать значения вероятностей для обоих классов, чтобы определить, к какому из них относится сообщение. После стадии обучения в классификатор передается тестовый массив данных и вычисляется результат. Точностью вычисления алгоритма является процент корректно определенных данных. Например, если в качестве входных данных подается 100 сообщений, которые гарантированно не являются спамом, то алгоритм должен показать точность равную 100 %. Для тестирования разработанных библиотек также применялась модификация алгоритма, учитывающая наличие часто повторяемых слов в обеих группах [2].

Алгоритм разработан на основе наивного байесовского классификатора, позволяющего определить вероятность события при условии, что произошло другое взаимозависимое событие: вероятность того, что сообщение принадлежит классу «не-спам»; вероятность того, что сообщение принадлежит классу «спам»; безусловную вероятность встретить сообщение класса «спам» в обучающей выборке сообщений; безусловная вероятность встретить сообщение класса «не-спам» в обучающей выборке сообщений.

Для тестирования алгоритма разработано программное обеспечение на мультипарадигмальном языке программирования *F#* из семейства языков *.NET Framework*, поддерживающем функциональное программирование в дополнение к императивному (процедурному) и объектно-ориентированному программированию, позволяющему писать абстрактные обобщенные алгоритмы, которые управляют параметризованными структурами данных, с использованием библиотеки *Accord.net*.

Разработанный алгоритм с применением наивного байесовского классификатора показал высокую эффективность и точность результатов.

При рассмотрении сообщений не только с точки зрения отдельных слов, но и их ассоциаций, рассматривается оптимизированный алгоритм, точность которого выше, чем у наивного байесовского классификатора. Таким образом, более полный анализ структуры текстовых сообщений позволяет не только повысить качество классификации, но и открывает большие перспективы с точки зрения разработки новых алгоритмов и оптимизации уже существующих.

Литература

1. Леонтьева, Н. Н. К теории автоматического понимания текста : в 3 ч. / Н. Н. Леонтьева // Изд-во Моск. ун-та. – 2002. – Ч. 3 : Семантический компонент. Локальный семантический анализ. – С. 49.
2. Коллекция смс спама. – Режим доступа: https://archive.ics.uci.edu/ml/datasets/sms+spam+collection. – Дата доступа: 12.10.2018.