УДК 004.912

МЕТОД ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА В ПРИЛОЖЕНИЯХ АНАЛИЗА ЭМОЦИОНАЛЬНОЙ ОКРАСКИ

Мамедова Э. Н. (Университет ИТМО, Санкт-Петербург)
Научный руководитель – к. т. н., Лапаев М. В.
(ООО "ТПП Лаб", Санкт-Петербург)

Работа посвящена методам обработки данных на естественном языке с целью анализа эмоциональной окраски для оптимизации бизнес-процессов. Проведен анализ существующих инструментов и условия их использования. Представлен легковесный алгоритм предварительной обработки текстовой информации, а также оценки его эмоциональной окраски с применением шаблонного поиска и словаря синонимов.

Введение. На сегодняшний день весомую долю в онлайн-коммуникациях занимают данные, представленные текстом. Данные, получаемые из блогов, отзывов, комментариев, форумов и прочих источников, где каждый желающий может выразить свое мнение, зачастую несут полезную информацию для маркетологов различных организаций, которые, в свою очередь, основываясь на анализе полученной информации могут корректировать стратегии развития и продаж своей продукции на основании мнения потребителя. Такая необходимая для построения бизнес-процессов информация неподвластна ручной обработке даже при наличии целого отдела сотрудников, поскольку потоки насчитывают тысячи сообщений в минуты. Компании прибегают к инструментам поиска и анализа текстовых данных, производители которых предоставляют услуги на условиях подписки с лимитированной пропускной способностью. Кроме того, применение сторонних сервисов повышает риск утечки данных, составляющих коммерческую тайну. Основной проблемой является вариативности на всех уровнях, поэтому метод, применимый для всех языков, не существует. Существующие алгоритмы либо узконаправлены и пригодны для решения задач в ограниченной предметной области, либо имеют широкое назначение и далеки от легковесных. Целью настоящего исследования является разработка и верификация легковесного метода предварительной обработки эмоционально-окрашенных текстов для выявления фрагментов, содержащих выражение эмоций, с целью последующей глубокой обработки посредством нейронной сети.

Основная часть. Анализ эмоциональной окраски текста - многоэтапный процесс, требующий упрощения текста на естественном языке до множества машинно-интерпретируемых токенов. На базе алгоритма существующего парсера SemSin с целью его улучшения был разработан оптимизированный алгоритм, позволяющий из абзаца сырого текста построить дерево зависимостей. На вход парсер принимает абзац текста, который затем передаётся в графематический анализатор, где разбивается на токены и предложения с применением набора простых регулярных выражений, а для разрешения неоднозначностей (сокращение/конец предложения и др.), применяются вероятностные модели, описывающие совместное распределение вероятностей меток токенов. Результатом работы графематического анализатора является разбитый на токены и предложения текст, который затем передаётся в морфологический анализатор,где каждому токену на основе словарей назначается набор характеристик, а для неизвестных словарю или редких слов для определения их характеристики применяются системы правил. При этом анализатор не всегда может выбрать единственный набор характеристик для токена в силу того, что на этом этапе пока ещё не разрешена частиречная омонимия. Для решения данной проблемы мы применяем вероятностные модели. После морфологического анализатора полученные токены с их характеристиками отправляются в классификатор, где каждому токену или последовательности токенов присваивается свой класс. В парсере SemSin для

классификации используются словари, регулярные выражения и наборы правил, но для повышения точности работы классификатора было решено со словарями и регулярными выражениями использовать классифицирующие вероятностные модели. Далее получившиеся наборы токенов с их характеристиками и классами отправляются в синтаксический анализатор, который строит дерево зависимостей. Полученное дерево зависимостей анализируется на наличие в последовательности эпитетов, сравнений, метафор и т.д., которые помечаются соответствующими маркерами. Для выявления эпитетов, метафор и сравнений применяется шаблонный поиск за счёт единой в общем случае структуры (шаблона) формирования данных литературных троп. Алгоритм шаблонного поиска реализован на основании правил, проверяющих последовательность слов на сходство с шаблоном того или иного тропа. Полученное множество выделенных фрагментов, носящих эмоциональную окраску, передаётся в модуль обработки на базе словаря синонимов, содержащего вероятности принадлежности слова к окраске трёх типов: отрицательная, нейтральная и положительная.

Выводы. Разработанный алгоритм предварительного анализа и разбора текста показал свою эффективность на массиве данных, поступающих со скоростью несколько сотен в минуту, и при этом не требует высокой производительности, что позволяет снизить затраты как на программные, так и на аппаратные средства.

Мамедова Э. Н. (автор) Подпись

Лапаев М. В. (научный руководитель) Подпись