УДК 004.912

ВЛИЯНИЕ ЭТАПОВ ПРЕДОБРАБОТКИ ТЕКСТА НА ЭФФЕКТИВНОСТЬ РАБОТЫ КЛАССИФИКАЦИИ МЕДИЦИНСКИХ ТЕКСТОВ

Кашина М. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., Копаница Г.Д.
(Университет ИТМО, г. Санкт-Петербург)

В связи с повсеместным введением электронных медицинских карт появилась необходимость в структуризации медицинских данных. В работе анализируется влияния предварительной обработки медицинского текста на эффективность работы классификатора.

Введение. Существует большое количество методов обработки текстов на английском языке, но при этом присутствует недостаток методов для других языков, в частности, для русского языка. Особенно для такой специфической предметной области, как медицина. Метод предварительной обработки данных играет важную роль и является первым шагом в обработке текста. Этап предобработки имеет решающее значение для определения качества классификации.

Основная часть. Цель работы: Оценить эффективность влияния этапов предобработки на дальнейшую классификацию аллергологических анамнезов. Данные были предоставлены ФГБУ «НМИЦ им. В. А. Алмазова».
Алгоритм работы включает в себя:

- поиск записей по ключевым словам: «аллергия» и «непереносимость»;
- удаление повторяющихся записей;
- удаление символов и лишних пробелов;
- исправление ошибок регулярными выражениями;
- обрезка документа;
- исправление ошибок расстоянием Дамерау-Левенштейна;
- нормализация;
- удаление стоп-слов;
- векторизация;
- гармонизация классов;
- классификация.

При обрезке документа самым эффективным оказался способ обрезки до 2 предложений и 10 значимых слов. Нормализация включала в себя перевод всех слов к одному регистру, приведение слов к начальной форме. Для векторизации использовался метод «мешка-слов» (Bag-of-words). Гармонизировали классы SMOTE алгоритмом. Для оценки эффективности предварительной обработки документов при классификации медицинских текстов сравниваются результаты точности, полноты и F-меры для обработанных и необработанных данных. В результате самыми эффективными этапами оказались обрезка, нормализация и исправление ошибок; удаление стоп-слов и SMOTE, наоборот, уменьшили значения метрик.

Выводы. Была проведена поэтапная предварительная обработка текста для оценки влияния каждого этапа на дальнейшую классификацию медицинского текста. Полученные результаты будут использованы в дальнейшем для извлечения аллергенов и реакций из медицинских записей.

Кашина М. (автор) Подпись

Копаница Г.Д. (научный руководитель) Подпись