УДК 004.934.2

ВЛИЯНИЕ НОРМАЛИЗАЦИИ ПРИЗНАКОВ ТЕКСТОВЫХ ТРАНСКРИПЦИЙ НА СЕНТИМЕНТ-АНАЛИЗ

Двойникова А.А., Карпов А.А.

(Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Университет ИТМО)

В работе рассматривается влияние нормализации признаков текстовых транскрипций речевых высказываний на сентимент-анализ данных. Экспериментальные исследования проводятся с использованием русскоязычной мультимодальной базы данных RAMAS. Она содержит в себе аудио и видео диалогов на различные тематики между двумя дикторами. Каждый диктор выражал одну из базовых эмоций: злость, страх, удивление, радость, отвращение, грусть, также в базе данных содержатся аудиозаписи нейтральной речи каждого диктора. Так как авторы базы данных не предоставили транскрипции речевых высказываний дикторов, то для экспериментальных исследований транскрипции извлекались с помощью автоматических систем распознавания речи, таких как SpeechKit от компании Яндекс и Speech Recognition от компании Google. Как было упомянуто выше, разметка базы данных происходила по 6 базовым эмоциям, для сентимент-анализа данных необходимо сгруппировать данные на 3 класса в соответствии с диаграммой Рассела: отрицательный класс включает в себя такие эмоции как злость, страх, отвращение, грусть, положительный класс – радость и удивление и нейтральный класс включает в себя нейтральную речь.

Сентимент-анализ текстовых данных происходит в соответствии со следующим алгоритмом: предобработка данных, векторизация данных и построение классификатора. Нормализация текстовых данных может происходить на двух этапах, на этапе предобработки и векторизации. Предобработка текстовых данных включает в себя токенизацию слов, понижение регистра, удаление пунктуации и стоп слов, и нормализацию слов. Нормализация слов может происходить с помощью двух методов: лемматизации (приведение слова к начальной форме) и стемминга (выделение основы слова). Для языка программирования Python существуют различные библиотеки для нормализации слова, которые работают отлично друг от друга. Так для лемматизации слов можно использовать следующие библиотеки: pymorphy2, pymystem3, natasha, для стемминга – RussianStemmer (nltk), SnowballStemmer (nltk), Stemmer. В работе показано влияние использования различных методов и различных библиотек для нормализации слов на сентимент-анализ. После этапа предобработки текстовых данных следует этап векторизации. Существует множество методов векторизации текстов, наиболее популярные из них это Word2Vec, FastText, BERT, ELMO. Наиболее распространенные методы нормализации векторов — это минимакс, z-масштабирование и десятичное масштабирование. В работе исследуется применение различных методов нормализации векторов в комбинации с различными методами векторизации текстов и их влияние на сентимент-анализ русскоязычных транскрипций речевых высказываний.

Двойникова А.А. (автор)