**Лингвистический анализа текстов социальных сетей**

***Мусатов Алексей Игоревич, Печененко Константин Сергеевич***

*Студенты*

*Алтайский государственный университет, Барнаул, Россия*

Работа посвящена результатам анализа текстов в социальных сетях, касающихся отношения к дистанционному обучению в период коронавируса. Авторы исследовали в социальных сетях, блогах, форумах и на других онлайн-площадках следующие проблемы: 1) удовлетворенность родителей, педагогов, школьников и студентов дистант-форматом обучения, 2) организационные, психологические, технические проблемы, с которыми столкнулись педагоги, обучающиеся и их родители, 3) положительные и негативные стороны дистанционного обучения. Посты и комментарии выгружались только из открытых источников с сохранением анонимности и конфиденциальности пользователей.

Основными задачами исследования являлись: изучение тональности мнений и позиций, а также построение моделей машинного обучения для автоматизации анализа текстовой информации. В процессе исследования определены основные тематики высказываний и осуществлена разметка текстовых сообщений по категориям. На основе анализа данных выявлены проблемные места в организационных, методических и технологических решениях, а также уровень психологической нагрузки для всех участников дистанционного обучения. В рамках исследования оценена успешность перехода общего, среднего специального и высшего образования на дистанционный формат.

Одним из результатов исследования является построение алгоритма классификации текстовых сообщений респондентов в группах социальной сети «ВКонтакте». Исходный объем данных представляет собой порядка 77 тысяч сообщений. Из них лишь около 7 тысяч высказываний были классифицированы при ручной разметке как относящиеся к дистанту. Для получения более сбалансированной выборки, пригодной для последующего построения классифицирующей модели, из всего объема данных было отобрано 11 тысяч сообщений из класса, не относящихся к дистанционному обучению. Каждое сообщение было токенизировано по словам с помощью библиотеки NLTK; все слова были лемматизированы (приведены к нормальным формам с помощью библиотеки pymorphy2; удалены стоп слова (слова, которые не несут смысловой нагрузки: предлоги, союзы и т.п.).

С помощью Count Vectorizer (библиотека Scikit-learn), преобразовывающего входной текст в матрицу, была получена матрица признаков X, значениями которой являются количества вхождения каждого слова в текст. Далее эти значения были взвешены с помощью TF-IDF. При этом вес каждого слова прямо пропорционален частоте употребления его в документе и обратно пропорционален частоте употребления этого слова во всех сообщениях выборки.

Целевой вектор *y* принимает значения: 0 – сообщение не относится к дистанционному обучению, 1 – сообщение относится к дистанционному обучению.

Для классификации использовалась реализация алгоритма градиентного бустинга Light GBM с максимальным количеством листьев дерева – 10, с максимальной глубиной дерева – 7, с количеством деревьев, равным 500. Для оценки качества модели использовались метрики: precision (точность), recall (полнота) и F-score. Значения метрик для тренировочной и тестовой выборок практически совпали, следовательно, построенную модель классификации можно использовать для автоматической разметки текстовых сообщений в социальной сети «ВКонтакте» на относящиеся к дистанционному обучению и не относящиеся, что и реализовано в ходе текущего исследования.