6. Гельгор А.Л., Попов Е.А. Технология LTE мобильной передачи данных: Учебное пособие. – СПб., 2011. –204 с.

УДК 004.67

Автоматическое выделение сообщений в социальных сетях, относящихся к дистанционному обучению в период короновируса

Козлов Д.Ю., Мусатов А.И., Хворова Л.А.
АлтГУ, г. Барнаул

В связи с переходом в марте 2020 г. всех учебных заведений на дистанционный режим обучения в средствах массовой информации и социальных сетях накопилось огромное количество информации, требующей определенной обработки, анализа и интерпретации. Перед исследователями встали задачи исследования следующих проблем: 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.

В рамках сотрудничества с лабораторией наук о больших данных и проблемах общества ТГУ возник совместный научный проект «Образование в условиях коронавируса: большие данные как инструмент измерения реакции общества». Стороны проекта изучают высказывания родителей, педагогов, школьников и студентов в социальных сетях, блогах, форумах и на других онлайн-площадках. Посты и комментарии выгружаются только из открытых источников с сохранением анонимности и конфиденциальности пользователей.

Группой ППС и студентов ИМИТ и МИЭМИС совместно с ТГУ определены основные тематики высказываний и осуществлена разметка текстовых сообщений по категориям. Одна из задач исследования – изучение тональности мнений и позиций, а также мониторинг информационного освещения темы дистанционного образования. Основной же задачей для участников проекта станет построение моделей машинного обучения для автоматизации анализа текстовой информации. Многие задачи станут темами выпускных квалификационных работ.

На основе анализа данных необходимо будет выявить проблемные места в организационных, методических и технологических решениях и снизить психологическую нагрузку для всех участников дистанци-

онного образования. Кроме того, необходимо оценить успешность перехода общего, среднего специального и высшего образования на дистанционный формат. Итогом проекта станет формирование рекомендаций по корректировке образовательной политики Министерства науки и высшего образования РФ.

В данной работе рассматривается решение одной из задач совместного проекта – построение алгоритма классификации текстовых сообщений респондентов в группах социальной сети «ВКонтакте» на сообщения, относящиеся и не относящиеся к дистанционному обучению с использованием языка программирования Python 3, который предполагается использовать для автоматической фильтрации сообщений для дальнейшего анализа.

Рисунок 1 – 100 самых частотных слов в сообщениях, относящихся к дистанционному обучению

Рисунок 2 – 100 самых частотных слов в сообщениях, не относящихся к дистанционному обучению

Набор данных для обучения представляет собой табличный файл с двумя колонками: первая содержит текстовое сообщение респондента; вторая – оценка эксперта относительно сообщения: относится ли это сообщение к дистанционному формату обучения или нет. Исходный объем данных – около 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 предобработано по следующей схеме: все

символы были приведены к нижнему регистру; буква «ё» была заменена на букву «е»; были удалены некириллические буквы и цифры. Далее сообщения были токенизированы по словам с помощью библиотеки NLTK; все слова были лемматизированы (приведены к нормальным формам: например, существительные приведены к единственному числу именительного падежа) с помощью библиотеки pymorphy2; были удалены стоп слова (слова, которые не несут смысловой нагрузки: предлоги, союзы и т. п.) [1].

На рисунках 1-2 представлены облака ста самых частотных слов в сообщениях респондентов, полученные с помощью библиотеки WordCloud. Размер слова в облаке прямо пропорционален его встречаемости.

Матрица признаков X была получена с помощью CountVectorizer из библиотеки Scikit-learn, преобразовывающего входной текст в матрицу, значениями которой являются количества вхождения данного 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 [2].

Целевой вектор y принимает значения 0 и 1 (0 – сообщение не относится к дистанционному обучению, 1 – сообщение относится к дистанционному обучению).

Данные были случайным образом разбиты на обучающую и тестовую выборки. Объем тестовой выборки составил 15% от всего объема данных.

Для классификации использовалась реализация алгоритма градиентного бустинга LightGBM [3] с максимальным количеством листьев дерева, равным 10, с максимальной глубиной дерева, равной 7, с количеством деревьев, равным 500. Для оценки качества модели использовались следующие метрики: precision (точность), recall (полнота) и F-score [4].

На тренировочной выборке значения метрик составили: precision = 0.898, recall = 0.894, F-score = 0.896. На тестовой выборке значения метрик составили: precision = 0.841, recall = 0.831, F-score = 0.835. Поскольку 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 планируется сделать в ходе дальнейшего исследования.

Библиографический список

1. Батура Т. В. Методы автоматической классификации текстов // Программные продукты и системы. – 2017. – №1. – С. 85-99.

2. Ramos, J. Using tf-idf to determine word relevance in document queries. // Proceedings of the first instructional conference on machine learning. – 2003. – Vol. 242. – С. 133-142.

3. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. Lightgbm. A highly efficient gradient boosting decision tree // Advances in neural information processing systems. – 2017. – С. 3146-3154.

4. Powers, D.M., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation // International Journal of Machine Learning Technology. – 2011. – Vol. 2. –С. 37-63.

УДК 004

Поиск способов реализации различных игровых стратегий в условиях неполной информации на основе нейронных сетей

Д.С. Козлов, О.Н. Половикова

АлтГУ, г. Барнаул

Существуют такие ситуации, когда игроки, в начале игры, владеют ограниченной информацией о стратегиях, а также о выигрышных функциях других игроков. В данных условиях, игрокам приходится принимать решения, которые основываются только на известной им информации. Подобные игры имеют своё название «игры с неполной информацией» либо «Байесовские игры». Стоит сразу отметить, что данного рода игры имеют существенное отличие от игр с несовершенной информацией, так как в таких условиях у игроков просто нет возможностей для наблюдения за действиями соперника.

Данный вид «неопределённости», зачастую связан с комбинаторными играми, в которых игроку предстоит сделать выбор из огромного числа стратегических вариантов. Соответственно, просчитать все возможные переборные варианты ходов «в уме» – не представляется возможным. И, в связи с этим, игроку в какой-то мере приходится выбирать «случайное» решение хода.

В настоящее время, выбор оптимального хода для игры с неполной информацией- возможен, например, при использовании нейронных сетей в данную игровую область нейронных сетей.