# Автоматические методы извлечения и кластеризации суждений по поводу коронавирусной инфекции из корпуса новостей и социальных сетей

***Нугаманов Эдуард Альбертович***

*Студент*

*Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия*

*E-mail:* ed.nugamanov@gmail.com

***Научный руководитель*** *— Лукашевич Наталья Валентиновна*

В настоящее время пандемия COVID-19 является одной из главных тем для обсуждения. Большие потоки информации об инфекции распространяются по социальным медиа [4]. Анализ такой информации мог бы принести практическую пользу. В частности, одним из возможных способов смягчить влияние пандемии на мир является изучение мнений в отношении самой инфекции и принимаемых ей в противодействие мер. Такой анализ мог бы выявить аспекты пандемии, вызывающие наиболее негативную реакцию у населения, и скорректировать политику борьбы с ней соответствующим образом.

Задача автоматического извлечения мнений — это широкое понятие, которое может включать в себя идентификацию, извлечение и анализ мнения автора в отношение чего-либо, выраженного в тексте. Цель данной работы – извлечение и кластеризация мнений пользователей по различным вопросам, связанным с COVID-19, например: масочный режим, вакцинирование, карантин. Для решения задачи были выполнены следующие шаги.

1. Был собран и обработан набор новостных статей о пандемии из социальной сети Telegram.

2. Полученный набор предложений был размечен на два класса: мнения и факты.

3. Были построены различные модели машинного обучения. Они обучались на составленной выборке, а их качество было оценено с использованием статистических метрик классификации методом кросс-валидации.

4. Полученный классификатор был применен к набору пользовательских комментариев из ресурса РИА Новости для улучшения выявления позиций пользователей по конкретным аспектам пандемии.

| | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| SVM+tfidf | 86.36 | 81.49 | 80.94 | 81.18 |
| SVM+fasttext | 84.67 | 76.19 | 84.37 | 80.05 |
| RuBERT | **89.77** | **86.79** | **84.59** | **85.67** |

Таблица 1: Сравнение моделей классификации

Для экспериментов использовались следующие алгоритмы машинного обучения: линейный SVM [2] с использованием tf-idf векторов документов, линейный SVM на fasttext [1] представлениях документов и нейросетевая модель на основе Sentence RuBERT [3]. Обучающая выборка составляла 1401 положительных и 2447 отрицательных примеров. В таблице 1 представлены значения основных метрик классификации, подсчитанных с помощью кросс-валидации на обучающей коллекции.

Подводя итог, в ходе данной работы удалось построить классификатор мнений, с помощью которого были проанализированы суждения пользователей о коронавирусной инфекции. Полученный результат может быть использован для кластеризации пользовательских высказываний, категоризации пользователей, а также для контроля распространения ложных фактов об инфекции.

**Литература**

1. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information // Transactions of the Association for Computational Linguistics, *5*, 2017, P.135–146, ISSN:2307-387X
2. Boser B., Guyon I., Vapnik V. A training algorithm for optimal margin classifiers // In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, USA, 1992
3. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // CoRR, *abs/1905.07213*, 2019, `http://arxiv.org/abs/1905.07213`.
4. Sharma K., Seo S., Meng C., Rambhatla S., Liu Y. COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations // arXiv:2003.12309 [cs.SI], 2020.