**Литературный обзор**

Анонимность в интернете иногда может заставить людей говорить гадости, которые они обычно не говорят в реальной жизни. В данной статье приведены примеры классификаций комментариев по определённым смысловым параметрам. Классификация комментариев проведена по двум классам - токсичные и нетоксичные. [1]

В данной литературе рассматривается модели - GRU и двунаправленный GRU, Двунаправленный LSTM и простой сверточный NN. Каждый твит о короновирусе проходит классификацию разными способами, после чего идёт сравнение и анализ данных моделей. Классификация твитов по коронавирусу проводится по позитивным, нейтральным и негативным оттенкам. [2]

Классификаторы логистической регрессии, реализованы в Apache Spark, т.е. на платформе для интенсивных вычислений в памяти. Основное внимание в статье уделяется сравнению этих классификаторов путем оценки точности классификации в зависимости от размера наборов обучающих данных и количества n-грамм. В экспериментах анализировались короткие тексты отзывов о товарах с сайта Amazon1. [3]

Предсказание оценки фильма. Датасет: набор данных состоит из файлов, разделенных вкладками, с фразами из набора данных Rotten Tomatoes. Разделение было сохранено для целей сравнительного анализа, но предложения были перетасованы по сравнению с их первоначальным порядком. Каждое предложение было разобрано на множество фраз анализатором Стэнфорда. У каждой фразы есть фразеид. У каждого предложения есть идентификатор предложения. Повторяющиеся фразы (например, короткие/распространенные слова) включаются в данные только один раз. [4]

В данной статье рассматривается классификация текста с примером отлавливания спама на почте. Обычной задачей в НЛП является классификация текстов или текстовых сообщений, это "классификация" в обычном смысле машинного обучения. Примеры включают обнаружение спама, анализ настроений и пометку запросов клиентов. [5]

В данной статье рассматривается классификация текстовых сообщений с помощью модели LSTM нейронной сети. Датасет содержит краткие сообщения пользователей. Модель учится разбивать сообщения на два класса - спам и не спам. [6]

С бурным развитием электронной коммерции в последнее десятилетие исследователи начали уделять больше внимания извлечению ценной информации из комментариев потребителей. В данной статье для получения семантических признаков предлагается метод тональности классификация на основе word2vec и SVMperf. Данные, используемые для обучения моделей, состоят из набора китайских комментариев о продуктах одежды. Результаты эксперимента показывают хорошую производительность представленного метода в классификации настроений. [7]

В данной статье проведен сравнительный анализ средств обработки текста. Рассматриваются разные способы математического представления текста, а также проводится классификация комментариев, основанная на разных представлениях текстовых данных. [8]

В статье проводится сравнительный анализ средств классификации текстовых данных. Рассматриваются такие модели, как: KNN, SVM, Random Forest, Boosting. Данные модели обучаются на выборке из банковских данных, для получения прогноза о выборе клиента открыть новый депозит. [9]

**Список источников**

1. <https://www.kaggle.com/jagangupta/stop-the-s-toxic-comments-eda>
2. <https://www.kaggle.com/maricinnamon/coronavirus-tweets-classification-nlp-gru>
3. <https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/5_2_05_Pranckevicius.pdf>
4. <https://www.kaggle.com/chiranjeevbit/movie-review-prediction>
5. <https://www.kaggle.com/matleonard/text-classification/notebook>
6. <https://www.kaggle.com/kredy10/simple-lstm-for-text-classification>
7. <https://sci-hub.ru/https://www.sciencedirect.com/science/article/pii/S0957417414005508>
8. <https://www.kaggle.com/adamschroeder/countvectorizer-tfidfvectorizer-predict-comments>
9. <https://www.kaggle.com/goldens/classification-review-with-python>