**Как работает программа?**

В ручном режиме - пользователь вставляет новость и нажимает проверить, размещенный текст обрабатывается. Эта обработка заключается в выявлении, из каких слов состоит текст, исключая группу очень распространенных слов (например: и, или, почему, поэтому…), знаков препинания и ссылок. Наконец, эти слова служат входными данными для 4 моделей, реализованных на платформе, которые будут указывать на вероятность фальшивых новостей.

В автоматизированном режиме – пользователь вносит ссылки на интересуемые ресурсы СМИ и запускает систему. По итогу получает последовательность значений «Правда!» или «Ложь!».

**Как обучалась модель?**

Модели, реализованные в системе, являются классификаторами, то есть возвращают вероятность того, что новость верна или ложна. Они изучают эти шаблоны из наборов входных данных и из новостных словарей.

Словарь состоит из более 100 000 новостей за последние 5 лет. Примерно половина из них правда и половина ложь.

Мы использовали базу данных, созданную NILC (исследовательская группа USP).

А также собственную выборку из новостных источников РФ, которые мы отметили, как правду.

**Какие параметры учитываются, чтобы считаться фейковой новостью или нет?**

Некоторые модели учитывают только частоту появления слов в тексте. На этапе обучения модель узнает, что некоторые фейковые новости используют, например, слова в повелительном тоне или с ощущением срочности, в то время как настоящие новости стараются быть более нейтральными.

Другие модели пытаются определить контекст, в котором встречаются слова, например, какие слова составляют этот текст и относится ли их набор больше к истинным или ложным новостям.

**В чём уникальность Вашей работы?**

Наша система изучает не только текстовые характеристики, такие как количество наречий, прилагательных, средний размер предложений, но и лексику, чтобы определить, правдивы новости или нет.

Кроме того, мы возвращаем непосредственный результат реализованных моделей с намерением обеспечить прозрачность для конечного пользователя.

Также наша система предлагает возможность работы не только с текстовой информации, но и с фото и видео. Эти модули работают на основе нейронной сети ResNet – 18.

**Как происходило обучение системы?**

Для обучения модели мы используем библиотеку GenSim, в которой уже реализованы некоторые методы машинного обучения для текстовых данных. Мы решили протестировать две реализации методологии: skip-gram и cbow.

В обоих случаях использовались следующие гиперпараметры.

**Какие модели были использованы?**

Ввиду этого мы решили скорректировать следующие модели.

* логистическая регрессия
* SVM (с использованием только ядра RBF)
* Случайный лес (RandomForestClassifier)
* Адабуст (AdaBoostClassifier)
* LightGBM (LGBMClassifier\*)

**Какие языки программирования и библиотеки были использованы?**

Языки программирования: Python и JAVA.

Фреймворк: Flask.

Библиотеки: OpenCV, Keras, TensorFlow, lxml, BeautifulSoup, Pandas, Request.