**1 слайд (титульный слайд)**

Здравствуйте уважаемая комиссия

Я Кондрашов Даниил Владиславович

Мой научный руководитель – доцент, кандидат физико-математических наук, Папшев Сергей Владимирович

Тема моей работы – автоматическая тематическая классификация новостного массива.

**2 слайд (постановка проблемы)**

Нахождение нужной информации сегодня является одной из важнейших задач. Однако полнотекстовый поиск по всем данным сразу часто оказывается неточным и неэффективным из-за слишком широкого круга источников и огромных объёмов просматриваемых данных. Поэтому критически важно сузить зону поиска. Одним из наиболее рациональных способ сделать это является тематическая фильтрация. Так выбрав только релевантную предметную область, мы резко повысим качество и скорость поиска.

**3 слайд (актуальность автоматического тематического моделирования)**

Однако ручная разметка имеет серьёзные ограничения. Во-первых, она субъективна, так как человек может ошибиться в определении темы или создать слишком узкие, необобщаемые категории. Во-вторых, она неэффективна, так как объёмы неразмеченных данных растут колоссальными темпами. Человек просто физически не успевает за этим потоком. Тогда необходим алгоритм автоматической тематической разметки.

**4 слайд (цель и задачи)**

В последнее время по популярности и качеству обработки текстовых данных лидируют нейронные сети, поэтому будет логично выбрать именно их для решения поставленной задачи.

Цель моей работы – разработка нейросетевого метода автоматической классификации новостей на основе тематической модели предметной области.

Ключевыми задачами для достижения поставленной цели являются:

1. Автоматический сбор новостей *(через парсинг);*
2. Их автоматическая разметка по темам с использованием тематической модели;
3. И обучение нейросетевого классификатора на этих данных.

**5 слайд (слайд с примером сайта)**

Данные были собраны с помощью парсинга с новостного сайта ВШЭ, который имеет следующий вид,

**6 слайд (слайд с примером данных)**

и помещены в excel-таблицу следующего вида.

**7 слайд (слайд с количественными характеристиками исходных данных)**

Всего было собрано 17 тысяч новостей, средняя длина новости составила порядка семиста токенов.

**8 слайд (слайд с количественными характеристиками обработанных и необработанных данных)**

Собранные данные были обработаны с применением различных методов. То есть: проведена токенизация, слова приведены к начальной форме, удалены неалфавитные токены, а также проведена фильтрация стоп-слов по словарю и с помощью TF-IDF метрики.

Разница количественных характеристик между подготовленными и неподготовленными данными представлена на слайде.

**9 слайд (слайд с графиком перплексии)**

Для разметки подготовленных данных было проведено тематическое моделирование с подобранными оптимальными гиперпараметрами. Для тематического моделирования использовалась библиотека BigARTM, для оптимизации гиперпараметров библиотека Optuna.

Как видно из графика перплексии модель сошлась.

**10 слайд (распределение новостей по темам)**

На основе построенной тематической модели каждому документу была присвоена наиболее вероятная тема. Получившееся распределение новостей по темам представлено на данном слайде.

**11 слайд (можно картинку с логотипами Hugging Face и Roberta + результаты обучения нейронной сети)**

На размеченных данных была обучена нейронная сеть-классификатор. Использовалась предобученная нейронная сеть-трансформер Roberta. Данный тип нейронной сети, как утверждается в литературе, наилучшим образом подходит для тематической классификации текстов. Веса модели были взяты с платформы Hugging Face, как с платформы с одним из наиболее удобных интерфейсов.

Качество дообученной на данных с различной подготовкой модели представлено на слайде. Видно, что точность обученной модели получилась ниже ожидаемой.

**12 слайд (результаты обучения нейронной сети + с разметкой от ВШЭ)**

Тогда был провён анализ качества вычисленной в ходе тематического моделирования разметки. Он показал, что полученные результаты связаны не с предложенным методом, а с:

1. Возможно, недостаточным объёмом данных, что могло повлиять на тематическое моделирование;
2. Или с недостаточной тщательностью тематического моделирования.

Косвенным подтверждением этому служит результат полученный при обучении той же нейросети на данных с ручной разметкой. Это показывает, что преложенный метод имеет потенциал при улучшении качества данных и при более тщательном тематическом моделировании.

**13 слайд (Вывод)**

Таким образом, в ходе работы были разработаны программные компоненты для сбора и подготовки данных, для разметки данных на основе тематического моделирования, а также для обучения и использования нейронной сети-кассификатора. Следовательно, все задачи были выполнены, цель достигнута.

**14 слайд (источники)**

**15 слайд (спасибо за внимание)**

Спасибо за внимание