МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра технологий программирования

ПОСТРОЕНИЕ СИСТЕМЫ АНАЛИЗА И СУММАРИЗАЦИИ ПОТОКА НОВОСТЕЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование
информационных систем
факультета КНиИТ
Рыданова Никиты Сергеевича

Научный руководитель
доцент _____ Б. А. Филиппов

Заведующий кафедрой
доцент, к. ф.-м. н. _____ И. А. Батраева

Саратов 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современном мире объём данных, размещаемых в сети Интернет, растёт с каждым годом. В связи с этим всё большее внимание уделяется инструментам, способным анализировать эти данные в автоматическом режиме. Это направление компьютерных наук принято называть «обработкой естественного языка» (Natural Language Processing, NLP).

Стоит отметить, что наука об обработке естественного языка развивалась ещё в 20-ом веке. Однако, если ранее существовавший инструментарий (как правило, основанный на статистических методах) позволял решать узкий класс задач с некоторыми ограничениями, то сейчас технологии глубокого обучения позволяют достаточно эффективно решать, на первый взгляд, нерешаемые задачи.

Зачастую атомарным элементом в решении конкретной задачи является отдельно взятый текст. Относительно его содержимого могут решаться классические для машинного обучения задачи, например, анализ тональности («настроения» текста) или его классификация.

Иногда отдельно взятый текст не представляет интереса, и его рассматривают среди множества других текстов и пытаются установить связи между ними, например, найти похожие тексты по заданному или сформировать кластеры текстов в зависимости от их семантической близости.

Среди множества семантически схожих текстов, можно выделить минимально необходимый объем информации, достаточный для понимания смысла текста, что называют суммаризацией. Это может быть нужно как в рамках оптимизации при применении алгоритмов обработки текстов, так и для оптимизации процесса их прочтения для человека.

Цель работы: построение системы анализа и суммаризации новостей.

В ходе работы должны быть решены следующие задачи:

- проанализировать актуальное состояние науки об обработке естественного языка;
- рассмотреть существующие подходы к анализу семантической близости текстов;
- рассмотреть существующие подходы к суммаризации текстов;
- провести сравнительный анализ методов кластеризации новостей;
- спроектировать архитектуру системы анализа и суммаризации потока но-

востей;

— разработать систему для сбора, анализа, суммаризации и визуализации пользовательских данных.

Методологические основы обработки естественного языка представлены в работах Х. Лейн «NLP в действии» [**?**], Я. Гудфеллоу «Глубокое обучение».

Теоретическая значимость исследования. Задача суммаризации новостей включает в себя широкий спектр инструментов для решения задач обработки естественного языка и машинного обучения. Качественное решение задачи сопряжено с необходимостью получать информативные векторные представления текстов на естественном языке, что является важной задачей в науке об обработке естественного языка с множеством практических приложений.

Практическая значимость исследования. Наблюдается значительный рост объема информации в сети Интернет, и людям всё сложнее фокусироваться на интересующей их информации. Существующие системы, как правило, предоставляют ограниченный набор общих рубрик из-за чего большая часть содержимого окажется нерелевантной для пользователя, не предоставляют пользователю возможности для добавления собственных источников, не учитывают исторический контекст при формировании контента. Разработка системы, учитывающей эти недостатки, позволила бы пользователю получать новостную повестку в структурированном виде.

Структура и объем работы. Магистерская работа состоит из введения, 5 глав, заключения, списка использованных источников и приложения. Общий объем работы — 75 страниц, из них 61 страница — основное содержание, включая 12 рисунков и 2 таблицы, цифровой носитель в качестве приложения, список использованных источников информации — 35 наименований.

1 Краткое содержание работы

1.1 Исследование предметной области

Раздел «Исследование предметной области» состоит из краткой справки о науке об обработке естественного языка, изучения влияния технологий обработки естественного языка на общество и месте науки об обработке естественного языка в качестве основы для задачи кластеризации новостей на сюжеты.

В тексте приводятся факты, позволяющие утверждать, что технологии обработки естественного языка будут оказывать значимое влияние на дальнейшее развитие общества. Акцент делается на массовом применении различных приложений науки, приводится статистика по использованию населением планеты социальных сетей, что в совокупности представляется доказательством практической значимости исследования.

Отдельно проводится анализ существующих решений: «Яндекс.Дзен», «Mail.ru Новости», «Рамблер/Новости» и другие. Отмечаются достоинства и недостатки этих решений. Исследуются подходы, которые зачастую используются в этих и аналогичных решениях. Обзор подходов сопровождается ссылками на актуальные статьи, такие как ...

1.2 Обработка естественного языка

В разделе «Обработка естественного языка» рассматривается минимальный теоретический материал, освоение которого необходимо для достижения поставленной цели.

1.2.1 Токенизация и предобработка

Выделяются следующие этапы обработки естественного языка для применения в практических приложениях:

- токенизация;
- предобработка текста (стемминг, лемматизация, удаление стоп-слов);
- получение векторного представления;

Отмечаются недостатки традиционного подхода к токенизации:

- потеря контекстуальной информации;
- плохая устойчивость к изменениям в языке;
- сильная зависимость от конкретной предметной области.

Указанные недостатки существенны, а потому предлагаются альтернативные подходы, которые используются в современных больших языковых мо-

делях — алгоритм Byte-Pair Encoding и алгоритм WordPiece. Эти алгоритмы позволяют избежать вышеописанных недостатков.

Нумерация полученных токенов позволяет получать базовые векторы, которые затем будут обработаны для получения информативных векторов, применимых для анализа семантической близости. Для получения информативных векторов рассматриваются TF-IDF («мешок слов») и Word2Vec.

1.2.2 Получение векторных представлений

Мешок слов (bag of words) — базовая модель представления текстов в виде числовых последовательностей. В этой модели каждый текст кодируется вектором размерности N, где каждая размерность вектора соответствует одному из терминов в исходном словаре.

Отмечаются важные свойства векторных представлений на основе модели мешка слов:

— размерность пространства зависит от размера словаря;
— не учитывают контекстную информацию, содержит статистические данные;
— формируют разреженное пространство;
— векторы различных терминов ортогональны.

Алгоритм Word2Vec противопоставляется как более сложный подход, позволяющий получить информативные векторные представления с помощью моделей нейронных сетей. Для этого нейронная сеть обучается на одной из двух задач: предсказания контекста по слову и предсказания слова по контексту.

Авторы Word2Vec породили не один метод, а целое семейство новых способов представления чисел с помощью обученных под конкретную задачу нейронных сетей. С этого момента появилось множество иных методов генерации векторного представления слов, основанных на нейронных сетях.

1.2.3 Обзор архитектур нейронных сетей

Среди возможных архитектур для обработки естественного языка выделяются следующие классы:

1. рекуррентные нейронные сети;
2. трансформеры.

Отмечаются основные недостатки рекуррентных нейронных сетей:

— рекуррентная структура затрудняет распараллеливание вычислений;

— по умолчанию рассматривает лишь предыдущие слова (возможны модификации с линейным повышением вычислительной сложности);
— внутреннее состояние модели плохо интерпретируется.

Указанные недостатки решает модель трансформера за счёт своей архитектуры. Её устройство подробно рассматривается в соответствующем разделе. Отдельно уделяется внимание BERT — модификации исходной архитектуры трансформера, которая в последние годы чаще используется для решения задач, связанных с поиском семантической близости.

1.2.4 Анализ семантической близости

Получение векторных представлений слов позволяет решать широкий спектр задач. Одной из таких задач может быть поиск семантической близких друг-другу текстов.

Предполагается, что хорошо подготовленные веса Word2Vec-подобных моделей распределяют слова в векторном пространстве согласно их смыслу, а расстояния между ними описывают то, насколько эти слова схожи или отличаются.

Обсуждается известный пример

$$king - man + woman = queen$$

и приводятся статьи, исследующие эту тему.

Наибольшее затруднение, однако, представляет задача чёткое определения того, что, собственно, является семантикой при решении задачи семантической близости.

В частности для одного и того же набора текстов возможна группировка по:

— эмоциональному тону предложения;
— фактическому и событийному содержанию;
— содержательному смыслу.

Традиционными метриками при анализе точек в пространстве являются:

— метрика Манхэттена;
— метрика Евклида;
— косинусное расстояние (или косинусная близость).

Для обработки естественного языка выбор метрики может существенно

влиять на результат. В частности, если векторное представление текстов получается из суммы векторных представлений отдельных компонентов, то длинные тексты могут иметь большую меру, чем короткие. Из этого следует, что применение метрики Манхэттена и Евклида может исказить результаты. Косинусное расстояние не обладает этим недостатком.

Наконец, описываются две основные формальные постановки задачи, применимые в рамках исследования — классификация и кластеризация. Отмечаются трудности в применении задачи классификации в данном практическом приложении, что вынуждает использовать алгоритмы кластеризации для решения задачи.

1.2.5 Задача суммаризации текстов

В рамках данного исследования сознательно не делается большого акцента на решение задачи суммаризации, поскольку задача кластеризации представляет собой более сложную, многоуровневую задачу. Кроме того, качественно сделанная суммаризация окажется напрасной работой, если новости сгруппированы неверно.

Выделяются два подхода к суммаризации текстов:

1. экстрактивная суммаризация (извлекает наиболее важные предложения из исходного текста);
2. абстрактная суммаризация, которая генерирует новые фразы и предложения на основе наиболее важных предложениях из исходного текста.

Приводится краткий сравнительный анализ этих двух подходов. Рассматриваются популярные предобученные модели, которые используются для решения задачи суммаризации. Отдельно отмечается тренд на применение больших языковых моделей, описывается инструмент Ollama, позволяющий легко применить такие модели на практике.

1.3 Обзор методов кластеризации

В этом разделе рассматриваются несколько методов кластеризации, которые могут быть применимы в рамках поставленной задачи.

Описывается алгоритм K-Means как базовый алгоритм для решения задачи кластеризации, затем рассматривается иерархическая кластеризация, изучается алгоритм DBSCAN, основанный на плотности точек.

Описание каждого алгоритма сопровождается их преимуществами и недостатками, сформированная теоретическая справка позволяет сделать вывод о хорошей применимости методов, основанных на плотности точек, так как эти методы позволяют исключать шумовые точки, которые неизбежно возникают в задаче анализа потока новостей.

При этом одной из наиболее критических проблем DBSCAN является его неспособность эффективно обрабатывать наборы данных с кластерами различной плотности. Поэтому далее подробно рассматривается модификация алгоритма DBSCAN под названием HDBSCAN.

Исследование алгоритма позволяет выделить следующие его свойства:

— учитывает локальную плотность точек за счёт метрики взаимной достижимости;
— строит иерархию, позволяющую получать кластеры разного уровня детализации;
— при правильной настройке устойчив к шуму;
— интуитивность настройки параметра `min_cluster_size`;
— вычислительно эффективный на большом наборе точек.

Применение алгоритмов кластеризации должно сопровождаться оценкой качества полученного разбиения. Для этого рассматриваются следующие метрики: коэффициент силуэта, индекс Дэвиса-Болдуина и индекс Calinski-Harabasz.

Отмечается, что внутренние метрики оценки качества кластеризации оценивают только её качество, но не качество на целевой задаче. Поэтому уделяется также внимание и внешней метрике оценки качества — Adjusted Rand Index.

Наконец, поскольку алгоритмы кластеризации чувствительны к размерности полученного векторного пространства, рассматриваются следующие алгоритмы для снижения размерности: метод главных компонент, алгоритмы t-SNE и UMAP. Выделяются особенности этих алгоритмов, сферы их применения.

1.4 Поиск оптимального алгоритма обработки данных

Данный раздел посвящен экспериментальной части работы. В нём проводится сравнительный анализ различных подходов к решению задачи.

Для проведения экспериментов вручную были собраны сюжеты из 850 новостей из Telegram-каналов «РИА Новости», «IF News» и «РБК. Новости. Главное» о геополитических и иных событиях в стране и мире. Общее число

сюжетов оказалось равно 71.

Кроме того, в набор данных было добавлено 50 настоящих новостей на случайные темы, не относящиеся ни к одному из существующих сюжетов, чтобы исследовать работу с шумовыми точками.

В рамках эксперимента исследовались несколько моделей векторных представлений:

— TF-IDF;
— предобученная модель Word2Vec (в составе библиотеки `FastText`);
— дообученная модель Word2Vec на 500 тыс. новостей из вышеуказанных каналов Telegram за последние 7 лет;
— `rubert-base-cased` в составе библиотеки `SentenceTransformers`;
— `jina-embeddings-v3` в составе библиотеки `SentenceTransformers`.

Такой выбор моделей связан с тем, что каждая из них представляет соответствующее поколение моделей с отличающимися друг от друга подходами.

Для TF-IDF данные проходили несколько шагов предобработки:

— удаление специальных символов (`r'[^a-яё\s]'`);
— приведение к нижнему регистру;
— лемматизация (`pymorphy3`);
— удаление стоп-слов (`nltk`).

При получении векторных представлений моделями на основе нейронных сетей использовались исходные текстовые данные без предварительной обработки.

Для кластеризации размерность все векторные представления были нормализованы, их размерность была уменьшена сперва до 100 с помощью метода главных компонент, а затем до 15 с помощью алгоритма UMAP с стандартными параметрами.

Применялось три алгоритма кластеризации — метод K-Means, HDBSCAN и агломеративная кластеризация.

Так как число кластеров заранее известно, то для оценки качества работы алгоритмов K-Means и агломеративной кластеризации использовалось именно это число кластеров. Понятно, что на практике эти значения будет необходимо подбирать на основе внутренних метрик кластеризации, что может дополнительно уменьшить точность работы алгоритма.

Для HDBSCAN число `min_cluster_size` также было выбрано на основе

априорных знаний о минимальном размере кластера в выборке. Понятно, что на практике это число должно выбираться на основе некоторых эвристик.

Для оценки результатов на целевой задаче использовался скорректированный индекс Рэнда. Кроме того, было решено переформулировать в задачу классификации: «Являются ли два текста частью одного сюжета?». Для этого были составлены уникальные пары точек с учётом их меток и определена точность, полнота и F-мера. Результаты исследования представлены в таблице 1.1.

		tf-idf	**tf-idf+svd**	**fasttext**	**fasttext-learned**	**rubert**	**jina-embeddings-v3**
kmeans	precision	**0.772657**	0.767469	0.505553	0.608243	0.346005	0.759490
	recall	**0.578416**	0.549911	0.349532	0.436205	0.245083	0.538761
	f1-score	**0.661574**	0.640726	0.413308	0.508055	0.286928	0.630361
	calinski_harabasz	2554.448486	2788.607422	925.233032	2311.368896	1131.205688	8015.352051
	silhouette	0.753295	**0.757708**	0.625595	0.628498	0.598846	0.682565
	davies_bouldin	0.598103	0.589810	**0.802394**	0.787381	0.823769	0.677768
	ari	**0.654624**	0.633536	0.401816	0.498205	0.272756	0.623008
hdbscan	precision	0.562774	0.651078	0.200233	0.381963	0.105149	**0.651093**
	recall	0.688861	0.730725	0.415904	0.615231	0.290207	**0.798885**
	f1-score	0.619467	0.688606	0.270322	0.471313	0.154367	**0.717457**
	calinski_harabasz	292.759033	477.133148	119.987114	150.920807	55.388271	742.226562
	silhouette	0.706061	**0.757537**	0.443326	0.576611	0.320351	0.749492
	davies_bouldin	0.984995	0.954637	1.134174	1.470679	1.215060	1.014606
	ari	0.609362	0.680669	0.246414	0.455525	0.124150	**0.709946**
agglomerative	precision	0.778601	**0.796107**	0.487777	0.603687	0.359038	0.775855
	recall	0.585569	0.572105	0.356790	0.444304	0.259072	**0.598927**
	f1-score	0.668428	0.665769	0.412126	0.511876	0.300971	**0.676006**
	calinski_harabasz	2694.059814	2969.301270	936.653992	2330.622559	1170.970825	8422.629883
	silhouette	**0.770700**	0.769109	0.626520	0.635622	0.600687	0.729992
	davies_bouldin	0.519210	0.571531	0.764794	0.751616	0.773085	0.576183
	ari	0.661601	0.659069	0.400217	0.501952	0.286924	**0.669233**

Таблица 1.1 – Показатели эффективности кластеризации для разных моделей

В процессе эксперимента также фиксировалось время, необходимое для вычисления векторных представлений на полученной выборке. Каждый метод запускался 20 раз с усреднением по времени.

Результаты измерений представлены на рис. 1.1. Для наглядности отображения по оси Y используется логарифмическая шкала.

Модель BERT показала наихудшие результаты с точки зрения внешних метрик, что неожиданно. Визуальный анализ продемонстрировал средний уровень качества, что может говорить о субъективности разметки и неоднозначности полученных результатов.

Важным результатом является то, что TF-IDF с применением сингулярного разложения незначительно уступает по метрикам, обеспечивая при этом значительное быстродействие, что может положительно сказаться на применении модели в практических приложениях.

Полученные результаты позволяют утверждать, что TF-IDF остаётся от-

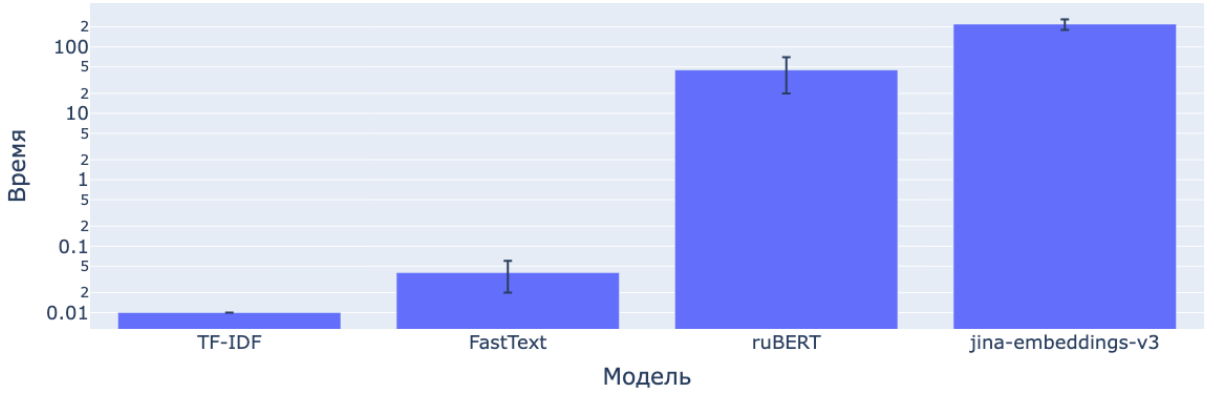

Рисунок 1.1 – Время, затраченное на получение векторных представлений в зависимости от выбранной модели

личным инструментом для поиска семантически близких новостей и может применяться в качестве базовой модели при решении различных прикладных задач в обработке естественного языка.

Важным результатом исследования является и то, что применение семантического моделирования действительно даёт прирост относительно статистических методов, что говорит о перспективности дальнейших исследований в этой области.

Таким образом, перспективной областью исследований может быть применение гибридных моделей из лексико-статистических методов и семантического моделирования в практических приложениях.

1.5 Практическая реализация системы

В заключительном разделе работы описывается практическая реализация системы. Формализуются требования к системе, описывается архитектура и её составные части.

Приводятся наиболее значимые фрагменты программного кода для сбора сообщений пользователя из Telegram, обработки полученных сообщений и отображения их конечному пользователю, демонстрируется интерфейс пользователя.

ЗАКЛЮЧЕНИЕ

В рамках данной работы была поставлена цель — построение системы анализа и суммаризации новостного потока. Для достижения цели были решены следующие задачи:

1. проведён анализ актуального состояния науки об обработке естественного языка, а также рассмотрены современные методы кластеризации и суммаризации текстов;
2. изучены и систематизированы существующие подходы к анализу семантической близости текстов;
3. рассмотрены основные подходы к суммаризации текстов, включая экстрактивные и абстрактивные методы, а также возможности их применения в задачах обработки новостных данных;
4. проведён сравнительный анализ методов кластеризации новостей: исследованы результаты работы алгоритмов K-means, агломеративной кластеризации и HDBSCAN на тестовых данных, что позволило выявить преимущества и ограничения каждого подхода для новостных сюжетов;
5. спроектирована архитектура системы анализа и суммаризации потока новостей, включающая сервисы сбора данных, обработки и кластеризации, а также визуализации результатов для пользователя;
6. разработана и реализована система, обеспечивающая автоматизированный сбор, анализ, кластеризацию, суммаризацию и визуализацию пользовательских данных, полученных из Telegram-каналов.

Таким образом, все поставленные задачи были успешно решены, а цель работы достигнута. В результате реализована система, позволяющая структурировать и сокращать большие потоки новостной информации, что подтверждается экспериментальными результатами и тестированием на реальных данных.

Среди дальнейших направлений развития работы можно выделить следующие:

- исследование методов обучения, позволяющих получить более информативные векторные представления;
- применение гибридных подходов, сочетающих простоту и стабильность TF-IDF и семантическое богатство векторных представлений, полученных с помощью нейронных сетей;
- проведение углубленного анализа алгоритма HDBSCAN,

— доработка и стабилизация прототипа системы;
— оптимизация времени работы системы за счёт инкрементальной кластеризации;
— более подробное изучение современных подходов к суммаризации текста.