Санкт-Петербургский государственный университет

Максимкин Матвей Сергеевич

Выпускная квалификационная работа

Методы интеллектуального анализа данных в задачах исследования пользовательского контента в социальных сетях

Уровень образования: бакалавриат

Направление 01.03.02 «Прикладная математика и информатика»

Основная образовательная программа СВ.5005 ««Прикладная математика. фундаментальная информатика и программирование»

Профиль «математическое и программное обеспечение вычислительных машин»

Научный руководитель:
доцент, кафедра технологий программирования
кандидат ф.-м. н. Сергеев Сергей Львович

Рецензент:
кандидат ф.-м. н. Кан Дмитрий Александрович

Санкт-Петербург
2022 г.

Содержание

Введение

В первом десятилетии 21 века распространение интернета, а также компьютеризация населения планеты вызвали информационную революцию. С развитием Веб 2.0 люди смогли начать делиться информацией друг с другом. А появление свободно доступных социальных медиа, таких как Facebook, Twitter, Instagram, Reddit привело к тому, что количество информации, создаваемой и распространяемой пользователями, колоссально возросло. Это количество быстро превысило те объемы, которые могут быть проанализированы классическими методами ручного труда исследователей и аналитиков. Однако, безусловная важность этой информации для задач бизнеса, ведения политики и обеспечения безопасности, стимулировало развитие компьютерных методов интеллектуального анализа данных в социальных сетях. Целью таких методов является выделение из больших массивов гетерогенной и неструктурированной информации, оставляемой пользователями, общих трендов, мнений или настроений, характеризующих отношение пользователей к разного рода вопросам, обсуждаемым в социальных сетях. Это необходимо для принятия дальнейших решений, объективно опирающихся на закономерности в общественном мнении какой-то конкретной предметной области.

Постановка задачи

Текстовые комментарии являются самой распространенной формой пользовательского контента. Каждый день пользователи оставляют миллионы комментариев и отзывов к публикациям в социальных сетях. В то же время социальные сети позволяют пользователям оценивать комментарии других людей с помощью механизма лайков, формируя рейтинг. На формирование этого рейтинга оказывает влияние множество факторов, некоторыми из них являются смысловое соответствие комментариев теме публикации, к которой они оставлены, а также тональность отраженных в них мнений. Целью этой

работы является анализ влияния этих двух показателей на формирование рейтинга комментариев в социальных сетях. В качестве объекта исследования взят один из самых популярных социальных медиа и седьмой сайт в мировом рейтинге по числу активной аудитории - Reddit.com. С помощью методов интеллектуального анализа для десяти самых активных сообществ интернет-портала разной тематики оценить релевантность пользовательских комментариев к темам соответствующих публикаций, а также их тональность. Исследовать зависимость семантической близости комментариев к теме публикации, а также их тональности по отношению к их рейтингу. Попытаться кластеризовать полученные данные с целью обобщить выводы о том, какой именно вид зависимости наиболее характерен для сообществ разной направленности.

Анализ литературы

Задача определения семантической близости между двумя текстовыми документами широко изучалась в области информационного поиска. Наиболее перспективным подходом является векторизация текстов. Процесс, при котором каждому документу в соответствие ставится вещественный вектор, отражающий его семантическую и лексическую структуру. Было разработано множество методов построения таких представлений. Простейшие основаны на представлении документов в виде взвешенных сумм, закодированных унитарным кодом, входящих в него слов [3]. Мера близости документов в таком представлении определяется пересечением входящих в них терминов. Другие используют модели машинного обучения для предварительного выявления синонимической связи между словами [7]-[8]. Недавние разработки в этой области привели к созданию моделей, способных учитывать семантику на уровне предложений, а также разделять слова, меняющие значения в зависимости от контекста [10]-[11]. Такие открытия позволили успешно применить методы анализа семантической близости к пользовательским сообщениям в социальных сетях, характерными чертами

которых является краткость и неоднозначность [9]. Однако, большинство работ в этой области были направлены на выявление комментариев, уводящих от темы в политических и новостных публикациях. Так, например, М. Мозафари и Н. Креспи использовали модели векторизации Word2vec и GloVe для детектирования комментариев, не связанных с темой, в новостных публикациях BBC на Facebook [6]. Насколько можно судить по открытым источникам, пока что не было работ, исследовавших зависимость изменения семантической близости комментариев к теме публикации от их положения в рейтинге.

Задача определения тональности состоит в определении эмоциональной оценки авторов по отношению к объектам, речь о которых идёт в тексте. Классическим подходом к решению может быть использование тезариусов тональности, словарей, сопоставляющих эмоционально окрашенным терминам категориальное или непрерывное значение тональности [1]. Основные значения таких словарей могут быть определены вручную, и впоследствии лексикон расширен с использованием онтологий как Wordnet [2]. Такие тезариусы могут быть успешно применены для анализа тональности комментариев в социальных сетях [5], однако поскольку тональности слов не являются постоянными, а зависят от контекста, ограничено предметной областью. Улучшить результат в задачах определения тональности может объединение методов векторизации слов с лексическим подходом [4].

Глава 1. Анализ Семантической близости

1.1 Сбор и подготовка данных

Комментарии в социальных сетях призваны быть средой, где пользователи обсуждают материал публикации, под которой они оставлены, однако некоторые пользователи игнорируют это правило и начинают обсуждать вопросы, не связанные с темой. Другие пользователи реагируют на это с помощью механизма лайков, продвигая или понижая место таких

комментариев в списке выдачи. Причина, по которой в качестве платформы для анализа этого эффекта взят именно Reddit в том, что система ранжирования на этом сайте отличается от большинства социальных медиа. Вместо типичных лайков и дизлайков реализована система голосов, каждый пользователь может повысить или понизить рейтинг другого комментария на 1 пункт. Такой механизм ранжировки позволяет однозначно сортировать комментарии в списке выдачи.

Структурно Reddit поделен на тематически ориентированные сообщества, называемые Subreddit. В этой работе были собраны данные из десяти популярных Subreddit разной тематики. Политической - r/Politics, новостной - r/News, r/Worldnews, образовательной – r/Science, r/Todayilearned, r/Changemyview и развлекательной r/Lifeprotips, r/AskReddit, r/Iaf, r/Askmen. Из каждого Subreddit для анализа взяты 100 самых популярных публикаций за последний год.

Как и в большинстве социальных сетей, комментарии на Reddit могут быть оставлены не только к самой публикации, но и к другим комментариям. Таким образом каждый пост можно представить в виде дерева, где каждый комментарий имеет родителя. Те комментарии, которые в качестве родителя имеют саму публикацию, назовем комментариями верхнего уровня. Будем рассматривать только эти комментарии, поскольку они напрямую относятся к теме публикации.

Текстовый контент на Reddit, как комментарии, так и текст публикаций представлен в виде специальной вариации Markdown формата, поэтому все служебные символы и фразы разметки были удалены. Из текста были также удалены слова на языках, отличных от английского и эмодзи. Комментарии, оставленные автоматическими ботами сообщества тоже, были удалены. На последнем этапе комментарии, содержащие менее двух слов, были удалены как недостаточно информативные. В конечном виде выборка содержит 1783032 комментариев и 1000 заголовков. Сбор данных осуществлялся с

помощью библиотеки praw, предназначенной для работы с Reddit-API. Текст заголовка объединялся с дополнительным контентом тела публикации, если таковой имелся.

Характерной особенностью Reddit является то, что пользователи не могут напрямую оставлять в комментариях медиафайлы, однако, могут прикреплять гиперссылки на медиафайлы, размещенные на других платформах. В собранном датасете имеется 22233 таких комментариев, что составляет 1.23%. Поскольку гиперссылки не могут быть корректно проанализированы методами интеллектуального анализа они были вырезаны из соответствующих комментариев. Однако стоит учесть, что это приводит к потере части информации. На рисунке 1 показано распределение комментариев, содержащих ссылки, ось абсцисс отражает место комментария в рейтинге соответствующей ему публикации - 100% соответствует самым непопулярным. Характерная равномерность распределения позволяет предположить о несущественном влиянии потери информации на последующие результаты.

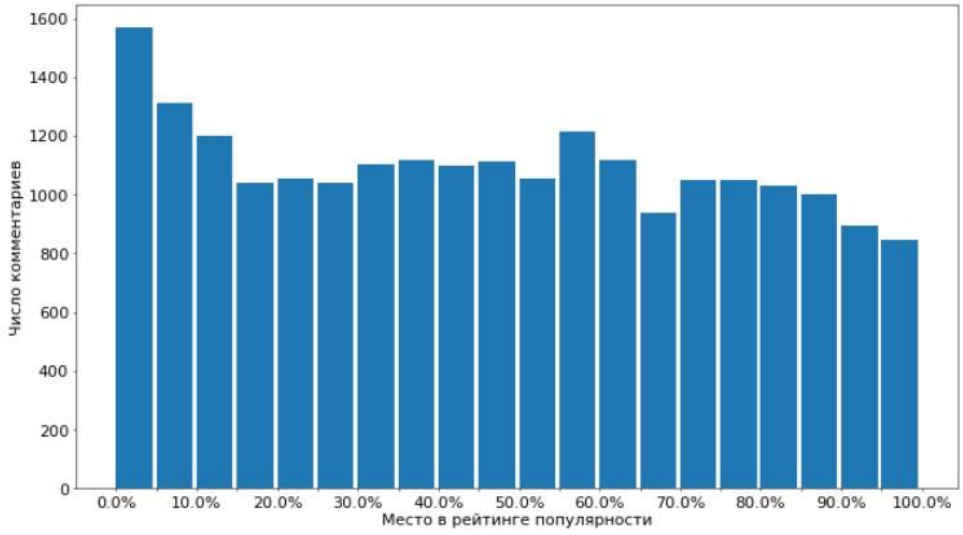

Рис. 1: Комментарии содержащие ссылки

1.2 Определение семантической близости

Ключевым шагом в задачах исследования естественного языка, в данном случае анализа смысловой близости пользовательских комментариев, является построение векторизации документов. Процесс, при котором каждому документу в соответствие ставится вещественный вектор, отражающий его семантическую и лексическую структуру. Классическим подходом является метод мешка слов. В этом методе каждое слово представляется в виде базисного вектора в пространстве размерности равной объему общего словаря документов. Затем представление каждого документа получается путем составления взвешенной суммы векторов входящих в него слов. В качестве весов могут использоваться частота слова в документе или же статистики, как tf-idf, отражающие важность каждого слова для этого документа в отношении к другим документам коллекции. Определение семантической близости двух документов закодированных таким образом сводится к вычислению числовой метрики, такой как евклидово расстояние или косинусовая мера между соответствующими векторами.

Однако такой подход имеет ряд недостатков. Высокая размерность векторов, при большом объеме словаря документов затрудняет вычисления. Каждый вектор слова ортогонален всем остальным, а значит не отражает лексическую связь между терминами. Также роли не играет положение слов в предложениях. Все вышеперечисленное делает данный метод непригодным для анализа пользовательских комментариев в социальных сетях, учитывая краткость таких сообщений, а также большой объем сленговой лексики.

Проблему определения синонимической связи между словами была решена с появлением моделей с использованием глубокого обучения, таких как word2vec, Glove и fastText. В основе этих моделей лежит статистическая гипотеза о том, что семантически связанные слова появляются в тексте в схожем контексте. При обучении на большом корпусе документов данные модели для каждого слова определяют небазисный вектор фиксированной

размерности, много меньшей, чем при использовании one hot encoding так, что слова, встречающиеся в тексте рядом с одинаковыми словами близки в смысле косинусовой меры.

Дальнейшие исследования в области анализа естественного языка привели к появлению моделей как ELMo, основанных на принципе двустороннего кодирования текстовых документов. Такие модели учитывают положение слов в предложениях, а также способны строить разные векторные представления для слов, меняющих свой смысл в зависимости от контекста. Последним качественным прорывом стало внедрение нейросетевых моделей на архитектуре трансформеров и появление модели Bert, на данный момент удерживающей лидерство в большинстве задач обработки естественного языка.

Предобученная модель Bert может быть использована и для анализа семантической близости текстов, однако в силу особенностей конструкции, для этого требуется производить попарное сравнение, подавая оба документа на вход нейросети. При достаточно большом размере выборки это приводит к колоссальному росту вычислительной сложности. Решением этой проблемы может быть извлечение независимых векторных представлений путем усреднения кодировок выходного слоя модели. Но представления, полученные этим методом, уступают более ранним алгоритмам векторизации. В 2019 году Н. Реймерс и И. Гуревич представили модель Sbert - модификацию Bert с использованием сиамских нейронных сетей, оптимизированную специально для решения задачи векторизации с целью определения семантической близости [12].

Для кодирования комментариев и текстов публикаций в данной работе использовалась модель «sbert-base-nli-mean-tokens» из библиотеки sentence_transformers. Затем для каждого комментария C и заголовка T была вычислена косинусовая мера сходства – формула (1), отражающая близость данного комментария к теме публикации.

$$cos\theta = \frac{C \cdot T}{\|C\| \cdot \|T\|} \quad (1)$$

Далее в каждой публикации комментарии были отсортированы по популярности в порядке убывания. Поскольку число комментариев в постах сильно разнится, они были разбиты на 20 равных интервалов. Для каждого интервала вычислено среднее значение семантической близости к заголовку. На рисунках 2-3 изображены две случайно выбранных публикации.

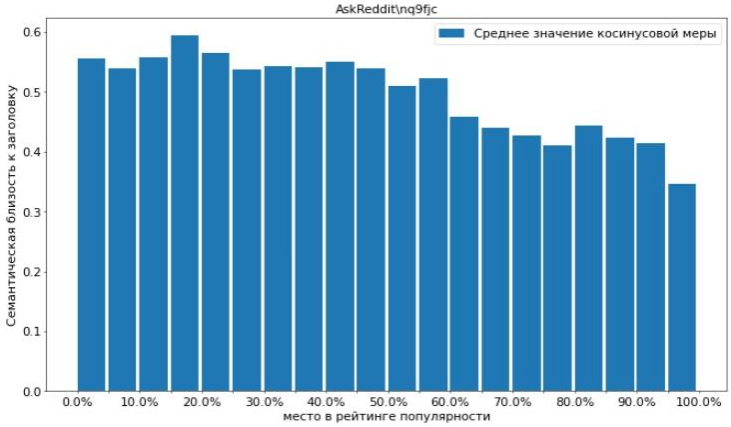

Рис. 2: Изменение семантической близости

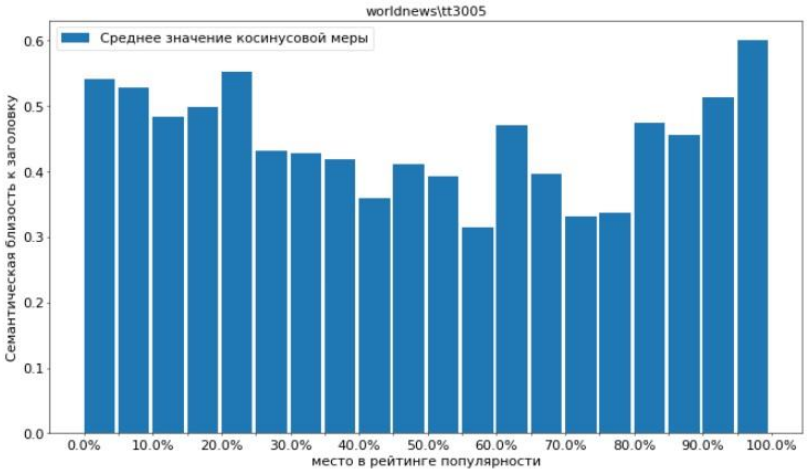

Рис. 3: Изменение семантической близости

1.3 Анализ зависимостей

Визуальный анализ множества графиков, полученных на предыдущем этапе, наводит на мысль, что динамика изменения смысловой близости комментариев к теме публикации с уменьшением популярности имеет

закономерности. Для их определения попробуем кластеризовать данные методом k-средних. Первым шагом проведем стандартизацию данных. То есть приведем их к виду, в котором среднее равно 0, а стандартное отклонение 1 – формула (2). Несмотря на то, что все значения семантической близости имеют единую размерность, диапазон изменения их величины может разнится даже в тех случаях, когда они предположительно отражают схожую зависимость. Это может быть связано с высокой вариативностью длины заголовка публикации. Некоторые имеют дополнительный прикрепленный текст, а другие состоят из одного короткого предложения, что делает невозможным достижение больших значений семантического сходства.

$$z_i = \frac{x_i - \overline{X}}{\sigma} \tag{2}$$

Алгоритм k-средних разбивает набор данных *X* на *k* наборов *S1,S2,…,Sk*, таким образом, чтобы минимизировать сумму квадратов расстояний от каждой точки кластера до его центра – формула (3). На первом этапе положение центроидов выбирается произвольно, далее каждой точке присваивается метка кластера с ближайшим центром. Новые центры кластеров вычисляются как центры масс сформированных групп. Процесс повторяется до тех пор, пока не выполнится одно из трех условий: изменение положений новых центров кластеров станет ниже заданного порога, новая итерация не будет приводить к изменению меток точек, заданных на предыдущем этапе, или не будет достигнуто максимальное число итераций. В данной работе использовалась реализация алгоритма k-средних из библиотеки sklearn.

$$\arg min_S \sum_{i=1}^{k} \sum_{x \epsilon S_i} \rho(x, \mu_i)^2 \tag{3}$$

Оптимальное количество кластеров будем искать с помощью метода локтя. Суть метода состоит в том, чтобы запустить алгоритм с разными значениями числа кластеров k. Для каждого k после завершения алгоритма

вычислить сумму квадратов ошибок, то есть квадратов расстояний от точек до соответствующих им кластеров. С увеличением k деление выборок будет становиться более точным и значение ошибки будет уменьшаться, однако в момент, когда k превысит число реально существующих кластеров, скорость ее уменьшения сильно уменьшится, что отразится на графике в виде изгиба – рисунок 4. Таким образом, будем считать k=3 оптимальным. При данном значении параметра векторы, соответствующие центрам кластеров, лучшим образом отображают виды зависимостей семантической близости комментариев к теме публикации по отношению к их популярности, рисунки 5-7.

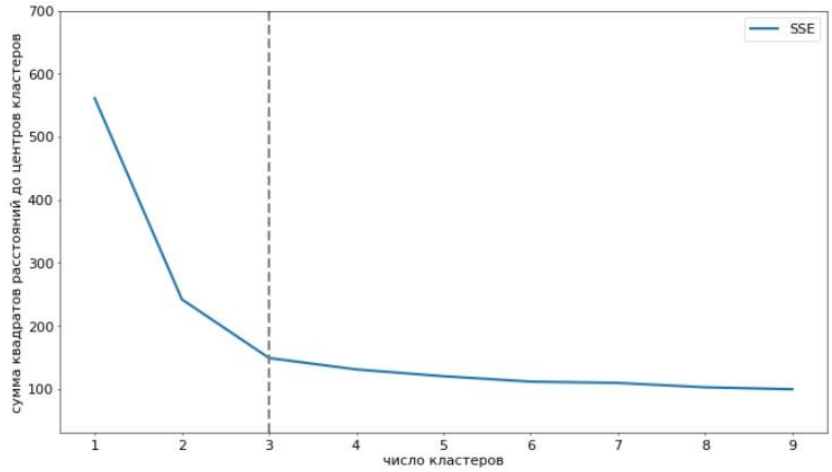

Рис. 4: Ошибка кластеризации

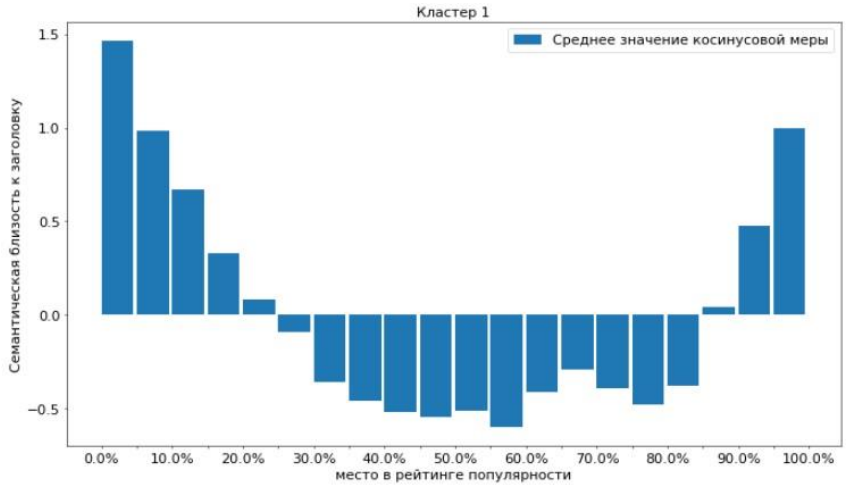

Рис. 5: Центр кластера 1

Зависимости, характерной для публикаций, отнесенных к первому кластеру, можно дать следующую интерпретацию. Пользователи, действительно, склонны ставить большинство голосов комментариям, больше всего соответствующим теме публикации. А при снижении популярности комментариев падает и смысловая связь с заголовком. Однако для комментариев, ниже всего оцененных пользователями, характерен резкий рост соответствия теме. Это может означать, что аудитория склонна сильнее всего отрицательно реагировать на комментарии, представляющие мнение, отличное от мнения большинства, чем на комментарии, не соответствующие теме.

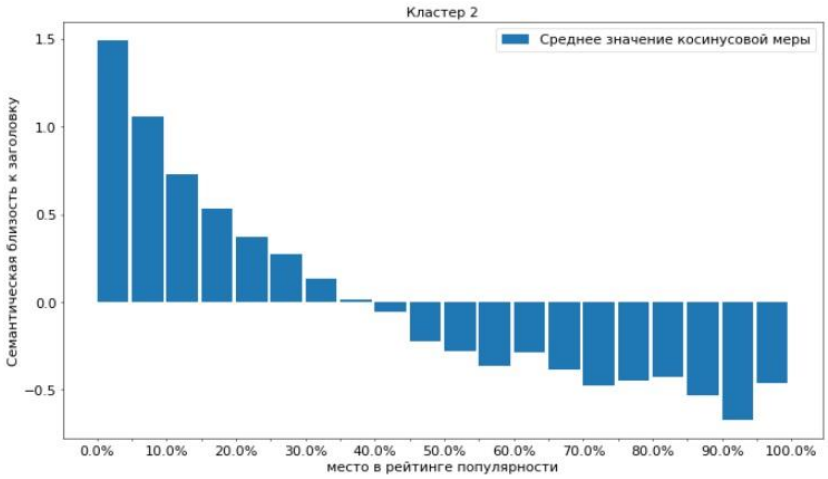

Рис. 6: Центр кластера 2

Второй тип зависимости можно описать как линейный тренд на снижение семантической близости комментариев и заголовка с уменьшением популярности. Чем больше авторы отходят от темы, тем меньше другие пользователи оценивают их контент.

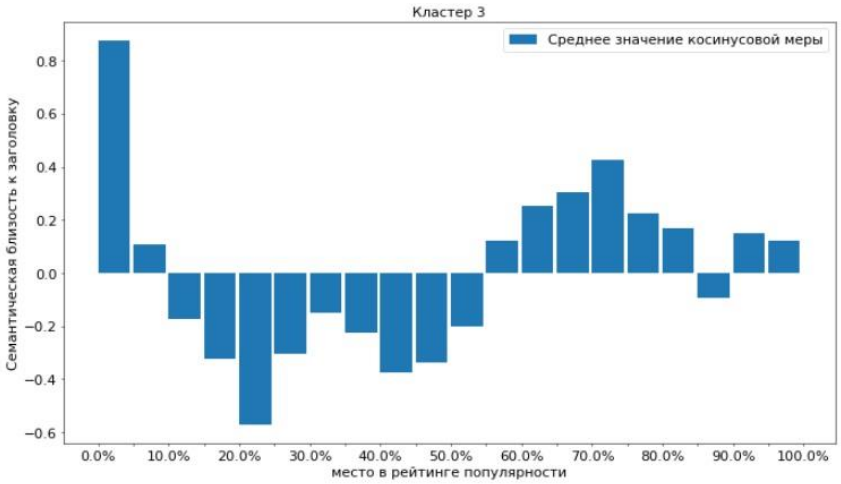

Рис. 7: Центр кластера 3

Для третьего типа зависимости характерен резкий спад смысловой связи с темой публикации и последующая ее вариация. Очевидное описание этой вариации трудно получить.

1.4 Анализ сообществ

В данной работе использованы данные из сообществ разной направленности. Поэтому для того, чтобы оценить влияет ли тематика Subreddit на характер изменения семантической близости комментариев к теме публикации, посмотрим на распределение публикаций, по соответствующим им по кластерам. Эта информация представлена в таблицах 1-2.

Таблица 1: Распределение публикаций по кластерам

	Subreddit				
	r/news	r/politics	r/worldnews	r/AskReddit	r/AskMen
Кластер 1	61	72	79	14	12
Кластер 2	28	23	13	35	36
Кластер 3	11	5	8	51	52

Таблица 2: Распределение публикаций по кластерам

	Subreddit

	r/ todayilearned	r/ science	r/ changemyview	r/ LifeProTips	r/ iaf
Кластер 1	15	24	19	26	17
Кластер 2	82	58	74	49	54
Кластер 3	3	18	7	25	29

Анализируя данные в таблицах, можно сделать вывод, что вид зависимости, соответствующий 1 кластеру, при котором самые непопулярные комментарии оказываются близки теме наиболее характерен для сообществ, связанных с политикой или современной повесткой: r/news, r/politics, r/worldnews. Причиной этому может служить высокая нетерпимость пользователей к альтернативному мнению по этим вопросам. Для большинства других сообществ самым распространенным является тип зависимости, соответствующий кластеру 2, при котором популярность комментариев уменьшается вместе с семантической близостью. Решение пользователей в среднем больше занижать оценки комментариям, уходящим от темы, можно считать естественным. Третий тип зависимости характерен только для сабреддитов с вопросами-ответами r/AskReddit и r/AskMen. Конкретное обоснование этому сложно получить, причиной может быть огромная разница между этими сообществами и всеми остальными в среднем числе комментариев у одной публикации. Количество комментариев сильно превышает те объемы, которые в среднем могут прочитать даже самые активные пользователи. Поэтому механизмы по которому комментарии ранжируются в таких публикациях могут отличаться.

Глава 2. Анализ тональности

2.1 Определение тональности

Для анализа тональности в данной задаче будем использовать лексическую модель Vader, основанную на словаре и разработанную

специально для анализа тональности в социальных сетях [5]. В основе Vader лежит словарь, сопоставляющий каждому слову или эмодзи эмпирически вычисленный индекс тональности в интервале (-4, 4), а также пять эвристических правил, модифицирующих это значение в соответствии с капитализацией слов, знаками препинания, частицами усиления или отрицания. Общее значение тональности для документа определяется путем нормализации суммы тональностей входящих в него слов - формула (4), где α – параметр, определяющий влияние длины документа на результат. Анализируя публикации в Twitter, авторы Vader определили α=15 как оптимальный. Учитывая, что средняя длина комментария в собранной коллекции составляет 25 слов, что на 7 больше чем в Twitter, будем использовать α=22.

$$\frac{x}{\sqrt{x^2 + \alpha}} \tag{4}$$

В данной работе использовалась реализация Vader из библиотеки nltk, подготовка данных осуществлялась по аналогии с главой 1, однако эмодзи из текста комментариев не удалялись. Отсортированные по популярности комментарии в каждой публикации, были также разбиты на 20 интервалов, для каждого интервала вычислялось среднее значение тональности. На рисунке 8 изображена случайная публикация из выборки.

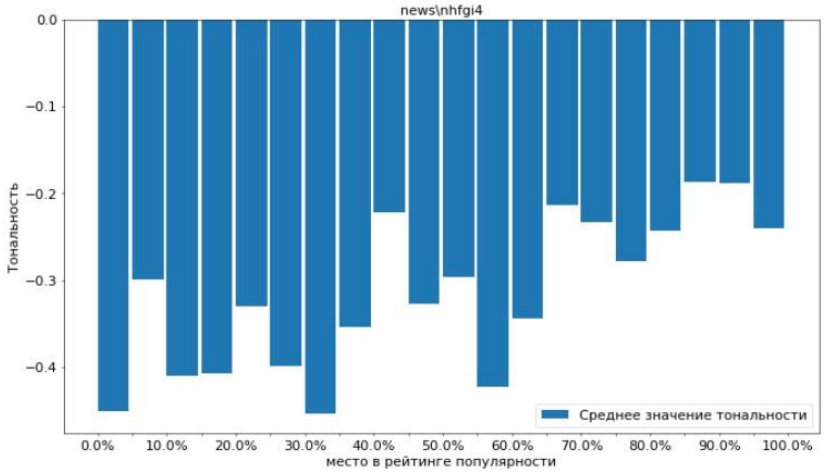

Рис. 8: Изменение тональности

2.2 Анализ зависимостей

Для анализа изменения тональности пользовательских комментариев по отношению к их популярности по аналогии с главой 1, попробуем кластеризовать данные методом k-средних. Перед началом стандартизируем данные по формуле (2). Оптимальное количество кластеров будем также искать методом локтя - рисунок 9.

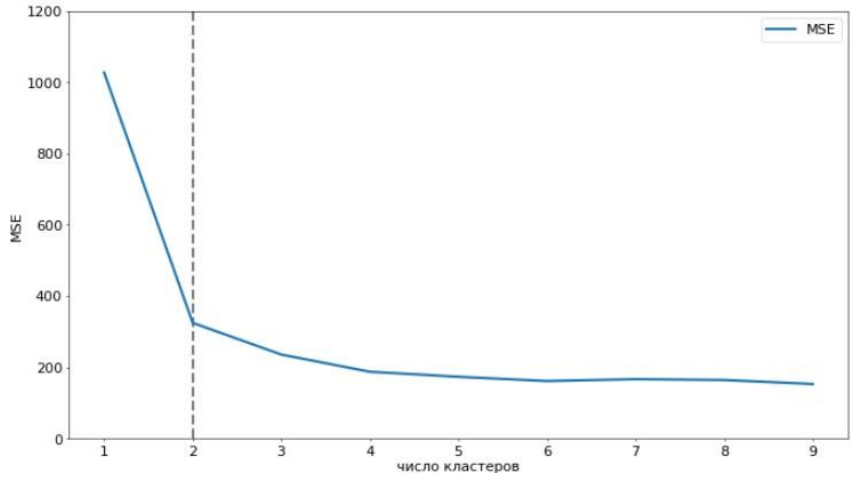

Рис. 9: Ошибка кластеризации

Будем считать, что k=2 лучшим образом характеризует данные. Для того чтобы оценить зависимость, построим графики векторов, соответствующих центрам кластеров - рисунки 10-11.

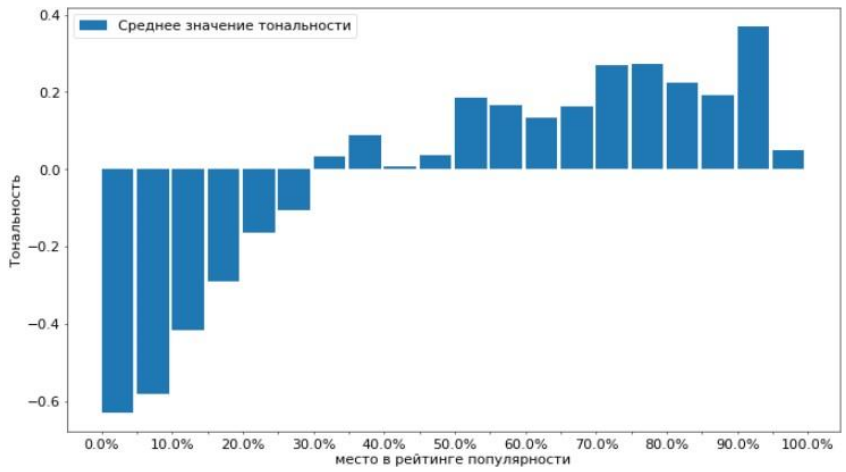

Рис. 10: Центр кластера 1

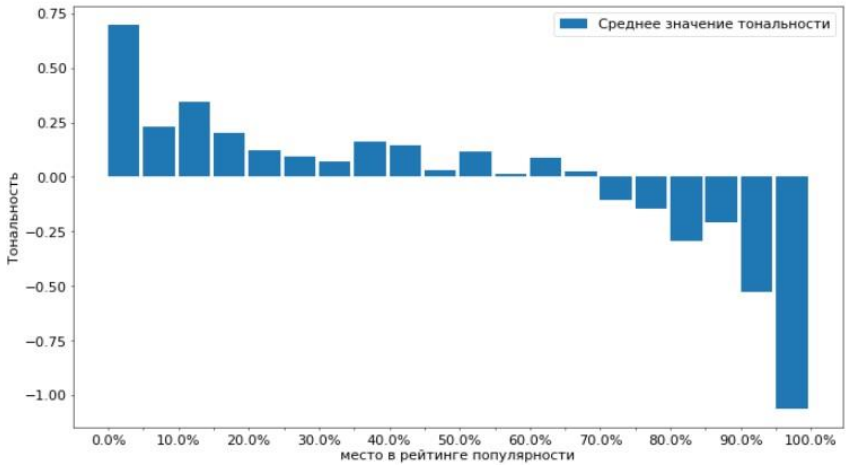

Рис. 11: Центр кластера 2

На графиках видно, что для комментариев с уменьшением популярности характерна постепенная смена тональности в противоположную сторону. Тем не менее важно отметить, что это не означает качественную смену тональности с позитивной на негативную или наоборот, а только динамику изменения. Также можно увидеть, что для обоих кластеров, для правого интервала, соответствующего 5% самых непопулярных комментариев характерно более резкое падение тональности. В таблицах 3-4 отражено распределение публикаций по кластерам для каждого Subreddit.

Таблица 3: Распределение публикаций по кластерам тональности

	Subreddit				
	r/news	r/politics	r/worldnews	r/AskReddit	r/AskMen
Кластер 1	76	71	58	34	37
Кластер 2	24	29	42	66	63

Таблица 4: Распределение публикаций по кластерам тональности

	Subreddit				
	r/ todayilearned	r/ science	r/ changemyview	r/ LifeProTips	r/ iaf
Кластер 1	19	44	41	61	55

Кластер 2	81	56	59	39	45

Первый тип зависимости, в которой тональность комментариев растет вместе с уменьшением популярности, характерен для сообществ политической и новостной направленности. Примечательно, что среднее значение тональности заголовков публикаций, отнесенных к этому кластеру, отрицательно и равно -0.122. Поскольку для анализа брались самые популярные посты за последний год, можно выдвинуть предположение, что в таких сообществах успех имеют публикации с отрицательным оттенком, а в комментариях пользователи негативно относятся к более позитивным мнениям по обсуждаемым вопросам.

Большинство публикаций, в которых понижение тональности комментариев коррелирует с популярностью соответствуют развлекательным сообществам r/AskReddit, r/todayilearned, r/AskMen. Средняя тональность заголовков в таких публикациях 0.09. Остальным сообществам не соответствует какой-то конкретный тип зависимости.

Выводы

Полученные результаты показывают, что семантическая близость к теме публикации, а также тональность действительно зависят от положения пользовательских комментариев в рейтинге. Однако необходимо сделать несколько критических утверждений. В частности, выбор количества интервалов для разбиения публикаций по популярности является произвольным, и может влиять на результаты. При выборе этого числа, нужно ориентироваться на размер самых маленьких публикаций в выборке. Оно должно быть в несколько раз меньше их объема, чтобы сглаживать влияние аномальных одиночных комментариев. Также на результаты влияет выбор метода векторизации текста и анализа тональности. В последнюю очередь хочется отметить, что на момент сбора данных часть комментариев была

удалена самими пользователями или модераторами. Таким образом информация в них содержащаяся не могла быть проанализирована. Они составляют 1.34% от общего размера выборки, и распределены неравномерно – рисунок 12, поэтому могли бы повлиять на результаты, если бы были доступны.

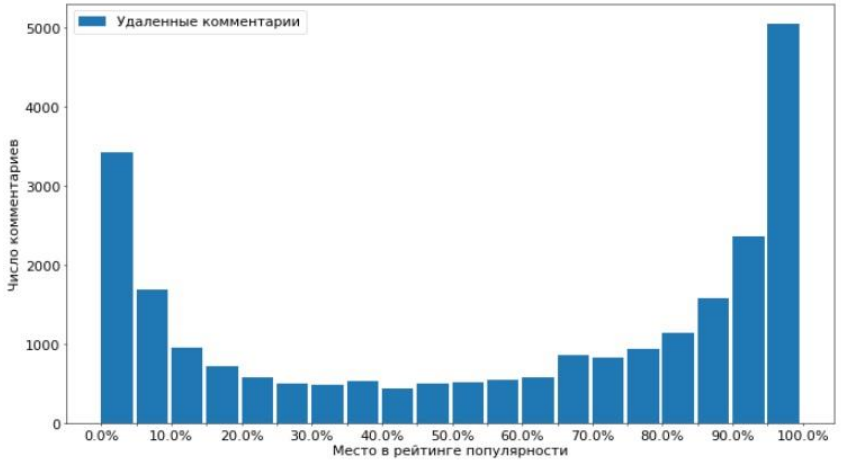

Рис. 12: Удаленные комментарии

Заключение

Проделанная работа показывает, что методы интеллектуального анализа данных могут успешно применяться с целью извлечения информации в задачах исследования пользовательского контента в социальных сетях. С помощью современных методов машинного анализа, можно выявлять закономерности из больших массивов открытых пользовательских данных в социальных медиа и с их помощью обосновывать теории описывающие общественные законы и поведение людей.

Список литературы

[1] *Baccianella S., Esuli A., Sebastiani F.* Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining LREC -2010. – Volume 10 P. 2200-2204

[2] Liu B., Zhang L. A survey of opinion mining and sentiment analysis Mining Text Data // Springer – 2012.- P. 416-417

[3] Qaiser S., Ali R., Use of TF-IDF to Examine the Relevance of Words to Documents // International Journal of Computer Applications – 2018. - Volume 181. - №1. – P. 25-29

[4] *Araque O., Zhu G., Iglesias C.* A semantic similarity-based perspective of affect lexicons for sentiment analysis // Knowledge-Based Systems – 2019. – Volume 165 P. 346-359

[5] *Hutto C., Gilbert E.* VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text // Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media – 2015. – P. 11-13

[6] Mozafari M., Farahbakhsh R., Crespi N. Content similarity analysis of written comments under posts in social media // 6th International Conference on Social Networks Analysis, Management and Security – 2019.- P. 158-165

[7] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space [Электронный ресурс] // arXiv.org. - URL: https://arxiv.org/abs/1301.3781 (дата обращения: 19.04.2022).

[8] *Bojanowski P., Grave E., Joulin A., T. Mikolov* Enriching Word Vectors with Subword Information [Электронный ресурс] // arXiv.org. – URL: https://arxiv.org/abs/1607.04606 (дата обращения: 19.04.2022).

[9] *Boom C., Canneyt S., Bohez S., Demeester T., Dhoedt B.* Learning Semantic Similarity for Very Short Texts [Электронный ресурс] // arXiv.org. - URL: https://arxiv.org/abs/1512.00765 (дата обращения: 01.05.2022).

[10] *Peters M., Neumann M., Iyyer M., Gardner M.* Deep contextualized word representations [Электронный ресурс] // arXiv.org. – URL: https://arxiv.org/abs/1802.05365 (дата обращения: 25.04.2022)

[11] *Devlin J., Chang M-W., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Электронный ресурс] //

arXiv.org. – URL: https://arxiv.org/abs/1810.04805 (дата обращения: 26.04.2022)

[12] *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс] // arXiv.org. – URL: https://arxiv.org/abs/1908.10084