Московский государственный университет имени М. В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Балакова Анна Сергеевна

Выявление поляризации мнений в новостных текстах методами обучения без учителя

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф.-м.н. - и.о. заведующего кафедрой

К.В. Воронцов

Москва, 2022

Содержание

1 Введение

В современном мире анализ текстов для выделения мнений принимает все большую популярность в исследованиях интеллектуального анализа текста и обработки естественного языка (NLP). Интернет площадки (социальные сети, интернет-магазины) позволяют пользователям выражать свое мнение на любую тему. Таким образом, выделение мнений, например, в отзывах товаров, является довольно популярной задачей. Мнения могут присутствовать в различных отзывах, комментариях, новостях и т.д.

В данной работе рассматривается задача выделения различных мнений именно из новостных текстов. Большинство политических событий довольно массово обсуждается и освещается различными издательствами. И очень часто средства массовой информации рассказывают о событиях, отражая произошедшее только с одной стороны. Это связанно с тем, что зачастую издательства зависимы от чего-либо, например, от географического положения, от политических взглядов руководителей. Поэтому читателю иногда известно о том или ином событии только с одной точки зрения. Поэтому задача разделения новостного потока на различные мнения является довольно важной в современном мире.

В данной работе стоит задача кластеризации потока новостей, относящихся к одному политическому событию, методами обучения без учителя (unsupervised learning) на различные мнения, которые выражают издательства. Такая задача называется анализом мнений (opinion mining). В данной работе будет рассматривать именно модель обучения без учителя, так как получить большую размеченную выборку по новостных текстах является довольно трудной задачей.

Одним из главных вопросов данной задачи является конкретизация того, что есть мнение. В данной работе предполагается, что мнение выражается через некоторую комбинацию синтаксических и семантических признаков текста. В результате данного исследования удалось построить модель, позволяющую с определенной точностью производить кластеризацию текстов на заранее известное число мнений.

2 Постановка задачи

Рассмотрим $\mathfrak{D} = \{D_i\}_{i=1}^{30}$ - некоторая коллекция текстовых документов. Каждый документ D_i представляет собой некоторую коллекцию новостных текстов, относящихся к одному политическому событию. Число новостей в каждом документе различно. Каждая новость в каждом документе относится к какому-то мнению. Число мнений в каждом документе различно, а также заранее известно для любого документа. Также в коллекциях могут присутствовать нейтральные мнения (мнения, не выражающее чью-либо позицию, а излагающие только факты).

В данной работе каждый новостной текст будет представляться в виде некоторой последовательности элементов различных типов, или модальностей (обозначается: M). В общем случае модальностями могут быть: слова или n-граммы слов, фото, жанры, категории, авторы, ссылки и т.д. Рассматриваемый способ решения основан на модели [1], где предполагается, что мнение выражается ввиде некоторой устойчивой комбинации определенных синтаксических и семантических признаков. В данной работе будут рассматриваться следующие модальности (признаки):

- субъекты и объекты текста из словаря W^{so},
- именованные сущности из словаря W^{ner},
- тонально окрашенные (положительно и отрицательно) слова из словаря W^{tonal}.

Для каждого документа необходимо построить тематическую модель так, чтобы в одном кластере (теме) оказались новости, выражающее одинаковое мнение. Рассмотрим отдельный документ $D \in \mathfrak{D}$. $W = \cup_{m \in M} W^m$ — множество (словарь) всех употребляемых терминов (элемент из W, который относится к одной из выделенных модальностей) в данном документе. Каждый новостной текст $d \in D$ представляет собой последовательность n_d терминов $(w_1, ..., w_{nd})$ из словаря W. Термин может встречаться в документе любое количество раз. Считается, что каждый термин w в каждом тексте d связан с некоторой темой $t \in T$, которая не известна. ($|T|$ число тем фиксировано для каждого документа).

Предполагается, что каждый документ представим ввиде множества троек (d, w, t), выбранных случайно и независимо из дискретного распределения $p(d, w, t)$, заданного на конечном множестве $D \times W \times T$. При этом в модели новостные тексты $d \in D$ и термины $w \in W$ являются наблюдаемыми переменными, а темы $t \in T$ являются скрытыми переменными. Основываясь на гипотезах: "мешка слов"(порядок слов в каждом тексте неважен) и "мешка документов"(порядок текстов в документе неважен) можно перейти к представлению слов текста в виде чисел n_{dw} - число вхождений термина w в документ d.

2.1 Вероятностная модель

Основываясь на [2] и [4], для того чтобы построить тематическую модель коллекции новостных текстов D — необходимо найти множество тем T, распределения $p(w|t)$ для всех тем $t \in T$ и распределения $p(t|d)$ для всех новостей $d \in D$. Основываясь на найденных распределениях в дальнейшем будет решаться задача кластеризации по числу мнений.

Основываясь на гипотезе условной независимости, которая предполагает, что появление слов в тексте d, относящихся к теме t, описывается общим для всей коллекции распределением p(w |t) и не зависит от данного d получим, что распределение представимо ввиде: $p(w|d,t) = p(w|t)$. Тогда имеем задачу:

$$p(w|d) = \sum_{t \in T} p(w,t|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}.$$

Таким образом задача тематического моделирования сводится к задаче поиска двух неизвестных матриц меньшего размера — матрицы "терминов-тем": Φ и матрицы "тем-документов": Θ.

Используем принцип максимума правдоподобия для поиска неизвестных параметров Φ, Θ:

$$p(D, \Phi, \Theta) = C \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi,\Theta}$$

Перейдем к логарифму правдоподобия и получим:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta},$$

$$\sum_{w \in W} \phi_{wt} = 1,$$

$$\sum_{tinT} \theta_{td} = 1$$

Заметим, что ограничения на ϕ и θ связаны с тем, что они должны быть корректными вероятностными распределениями. Данная задача решается с помощью EM-алгоритма.

Регуляризация Задача тематического моделирования является некорректно поставленной, так как существует бесконечно много ее решений. Это в свою очередь ведет к неустойчивости модели, так как при различных начальных приближениях будут получаться различные результаты. Для решения этой проблемы в общем случае используется подход, основанный на регуляризаторах. Он заключается в том, чтобы некоторым образом ввести ограничения на Φ, Θ, тем самым уменьшить множество решений. Библиотека bigartm, которая используется в построенной модели, позволяет добавлять с определенными весами разные регуляризаторы, которые вполне интепретируемы.

В предположениях тематической модели каждое слово порождается определенной темой. При этом существуют слова, имеющие высокую вероятность принадлежности к разным темам, они называются фоном. Фон - это слова общей лексики, стоп-слова. Таким образом, в нашей модели есть как предметные темы, которые выражают мнение автора, так и фоновые. Для устойчивости результатов введем дополнительно тему, отвечающую словам общей лексики. Таким образом, будет решаться задача кластеризации, при этом одно из моделируемых мнений будет считаться фоновым.

Считается, что предметные темы выражаются небольшим количеством термов, которые называются ядром темы (в нашем случае - мнения). Данное предположение основано на гипотезе разреженности: каждый документ d и каждый термин w связан с небольшим числом тем t. А это значит, что большая часть вероятностей p(t| d) и p(w |t) должна быть близка к нулю. Так как если термин относится к большому числу тем, то, скорее всего, это стоп-слово, которое не имеет большую роль при определении тематики документа. То есть искомые распределения тем должны быть разреженными. Поэтому в модель добавляется разреживающий регуляризатор для предметных тем. Данный регуляризотор имеет вид:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_w \ln \theta_{td}$$

Фоновые темы, наоборот, должны быть как можно больше похожи на равномерное распределение, что отвечает сглаживающему регуляризатору, который имеет вид:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_w \ln \theta_{td}$$

Помимо сглаживающих и разреживающих регуляризаторов будем рассматривать также декоррелирующий регуляризатор для матрицы Φ. Его добавление основано на том, что различные темы должны быть как можно меньше похожими друг на друга и среди них не было одинаковых:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in Tt} \sum_{w \in W} \phi_{wt} \phi_{ws}$$

Учитывая введенные регуляризаторы перейдем к новой тематической модели, которая имеет вид:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

$$\sum_{w \in W} \phi_{wt} = 1,$$

$$\sum_{tinT} \theta_{td} = 1$$

2.2 Синтаксические признаки

Для выделения субъектов и объектов из текстов используются синтаксические деревья. Для получения данной структуры необходимо использовать методы, основанные именно на русском языке, так как синтаксические структуры в разных языках отличаются. В данной работе синтаксические деревья будут строиться с использованием библиотеки natasha, которая позволяет проводить глубокий анализ текстов именно на русском языке. Отметим, что она основана именно на новостных текстах. Пример разобранного предложения представлен на Рис. 1.

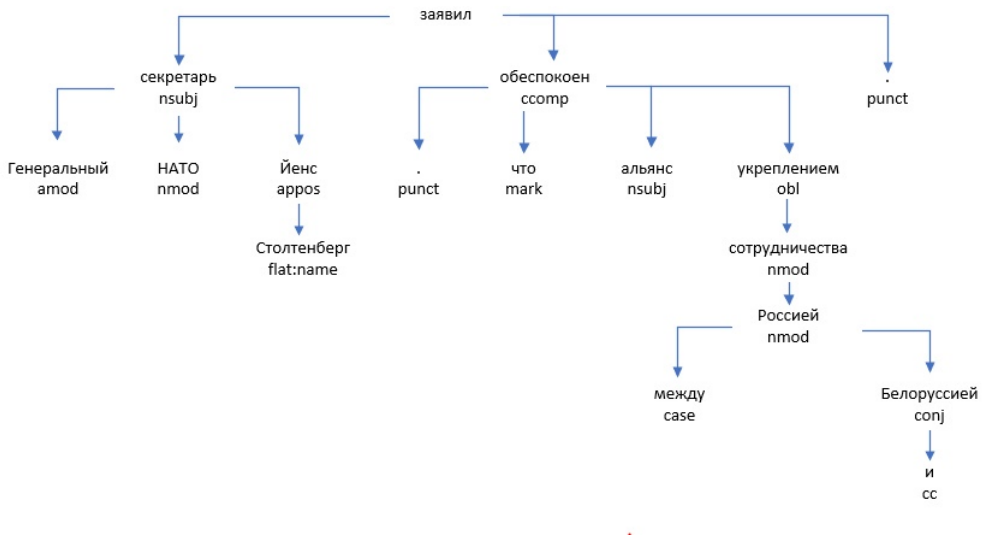

Рис. 1. Пример синтаксического дерева

Основываясь на синтакисическом разборе предложения можно выделять различные признаки. В данной работе в качестве признаков рассматриваются субъекты (nsubj) и объекты (obj). Данное предположение основано на том, что автор текста может выражать свою позицию рассказывая новость, то есть некоторые факты от определенного лица. И представление текста не в виде мешка слов (bag of words), а в виде мешка субъектов или объектов позволяет улучшить качество моделей в различных задачах.

2.3 Тональные слова

Выделение поляризованных мнений с помощью тонально окрашенных слов является довольно популярным методом. Он основан на том, что слова имеют определенное настроение и проанализировав распределение таких слов можно сделать вывод о том, что текст относится к определенному мнению. Для выделения таких слов существуют различные методы, которые позволяют выделять различные сентименты из текста. Настроения бывают разных видов, в данной работе будут рассматриваться только положительно и отрицательно окрашенные слова.

Тональные слова будут выделяться с использование словаря тональных слов (КартаСловСент [3]). Такой способ является довольно простым и не требует сложного анализа предложений: необходимо проверить наличие слов предложения в данном словаре. Но использование только слов, полученных таким методом, приводит к довольно скудному словарю признака, так как из небольшого новостного текста выделяется не так много тонально окрашенных слов. Для расширения данного признака будем использовать следующие методы, основанные на морфологическом анализе слов:

- если выделенное слово является существительным (NOUN) или прилагательным (ADJ), то родитель этого слова также помечается такой же тональностью, а также в признак добавляется пара из этого слова и его родителя,
- если выделенное слово является глаголом (VERB), то все связанные с ним субъекты и объекты помечаются той же тональностью, и в признак также добавляется пара из данных слов.

Также стоит учитывать, что в тексте могут встречаться отрицания, которые меняют сентимент связанного с ним слова. Поэтому при выделении нового тонально окрашенного слова проверяются все связанные с ним сущности. Если находится отрицательная частица, то сентимент слова меняется на противоположный.

Пример работы алгоритма приведен на Рис. 2

Генсек НАТО испугался дружбы России и Белоруссии: Альянс готовится к защите. В НАТО испугались дружбы России и Белоруссии. Как заявил генеральный секретарь альянса Йенс Столтенберг, там обеспокоены растущим сотрудничеством наших двух стран. В НАТО заявили, что готовы оборонять своих союзников от угроз, которые исходят от России или Белоруссии. "НАТО - оборонительный альянс. Я не хочу слишком много спекулировать. Но мы бдительны и очень внимательно следим за тем, что происходит в Белоруссии. Конечно, мы готовы в случае необходимости защищать и оборонять каждого союзника от любой угрозы, исходящей от Минска и Москвы", - заявил глава альянса в интервью изданию Welt am Sonntag. Столтенберг напомнил об основном подходе к отношениям с Москвой. Он описал его словами "сдерживание и диалог".

Рис. 2. Пример работы алгоритма выделения тональных слов

2.4 Именованные сущности

Задача выделения именованных сущностей (Named-entity recognition, NER) в задаче обработке естественного языка является довольно популярной. Цель данной задачи - выделить спаны (непрерывные части текста), в которых содержатся определенные сущности (в данной задаче рассматриваются персоны (PER), локации (LOC) и организации (ORG)). Пример выделения таких сущностей приведен на Рис. 3. Для решения данной задачи также используется библиотека natasha, позволяющая качественно и быстро выделить в тексте спаны.

Предполагается, что выделение именованных сущностей и их типов позволит улучшить качество существующей модели, основанной на триплетах субъект-предикат-объект (SPO), тонально окрашенных словах и семантических ролях. Это связано с тем, что автор рассказывает о произошедшем событии с точки зрения одной из сторон. То есть использует определенные именованные сущности (имена людей, названия организаций) с разной частотой.

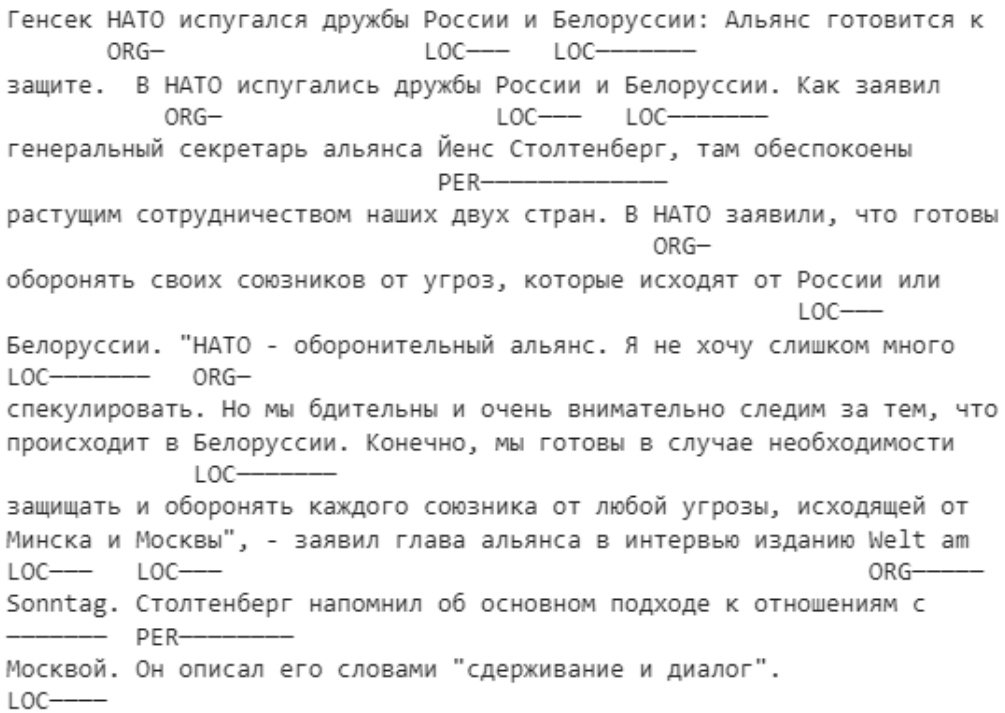

Рис. 3. Пример работы алгоритма распознавания именованных сущностей

2.5 Кластеризация

2.5.1 Метод K-средних

Данный метод кластеризации является довольно простым в реализации хоть и не очень точным. Он разбивает элементы выборки на заранее известное число кластеров. Алгоритм стремится минимизировать среднеквадратичное отклонение на

точках каждого кластера. Основная идея - на каждой итерации перевычисляется центр масс каждого кластера, полученного на предыдущем шаге, затем вновь производится разбиение. Алгоритм завершается, когда перевычисление центров масс не изменяет вид кластеров.

Данная модель кластеризации в данной работе рассматривается как некоторая начальная модель. Сравнение с ней позволяет доказать разумность использования более сложной модели кластеризации - вероятностного тематического моделирования.

2.5.2 Вероятностное тематическое моделирование

Данный метод кластеризации основан на построенной ранее тематической модели. Результатом тематического моделирования являются матрицы:

- $\Phi = \{\phi_{wt}\}, \phi_{wt} = p(w|t)$ - матрица "терминов-тем". То есть элементами матрицы являются вероятности принадлежности слова w к теме t,
- $\Theta = \{\theta_{td}\}, \theta_{td} = p(t|d)$ - матрица "тем-документов".

Отметим, что каждый столбец матрицы Θ содержит распределение мнений в документе $d : \theta_d = \{\theta_{td}, t \in T\}$. И основываясь на этих данных будем проводить мягкую кластеризацию тектов по мнениям.

Для каждого текста мы имеем вектор, состоящий из вероятностей принадлежности данного текста к каждой теме, то есть мнению. Будем относить текст к одному из мнений по максимальной вероятности данного вектора, т.е. $y_d = \underset{t \in T}{argmax}\ \theta_{td}$.

2.6 Метрики качества

Оценить качество кластеризации можно несколькими способами. Существуют следующие виды метрик:

- внешние меры: основаны на сравнении результата кластеризации с априори известным разделением на классы,
- внутренние меры: отображают качество кластеризации только по информации в данных.

Будем рассматривать внешние метрики, так как нам заранее известно правильное разбиение по кластерам. Данное разбиение получено с помощью разметчиков, которые проводили анализ новостных текстов. Таким образом у нас имеется эталонное разбиение на классы $T = \{t_1, ..., t_n\}$ и разбиение, полученное с помощью нашей модели $C = \{c_1, ..., c_m\}$.

При выборе метрики стоит учесть, что выбранная метрика должна удовлетворять некоторым условиям. В статье [5] приведен ряд принципов, которым метрики должны

удовлетворять. Также в этой статье оценивается их корректность и то, как различные типы метрик удовлетворяют им. Эти принципы:

- однородность кластеров (cluster homogeneity) - значение метрики качества должно уменьшаться при объединении в один кластер двух эталонных,
- полнота кластеров (cluster completeness) - значение метрики качества должно уменьшаться при разделении эталонного кластера на части (двойственное свойство к однородности),
- rag bag - при отнесении нерелевантного всем кластерам объекта значение метрики качества должно быть выше у той модели кластеризации, которая помещает данный элемент в шумный кластер, чем у модели, которая помещает этот элемент в чистый кластер,
- size vs quantity - ухудшение кластеризации большого числа небольших кластеров должно обходиться дороже небольшого ухудшения кластеризации в крупном кластере.

Рассмотрим BCubed-метрики, а именно точность, полноту и F-меру. Пусть имеется D - множество объектов, требующих кластеризацию. Тогда $d \in D$ - отдельный элемент этого множества, в нашем случае - новость, которую необходимо отнести к одному из мнений. Тогда $t(d)$ - эталонный кластер, которому принадлежит объект d, а $c(d)$ - кластер, к которому относит построенная модель. Тогда:

$$BCP = \frac{1}{N}\sum_{d\in D}\frac{|t(d)\cup c(d)|}{c(d)}, \quad BCR = \frac{1}{N}\sum_{d\in D}\frac{|t(d)\cup c(d)|}{t(d)}, \quad BCF = \frac{2\cdot BCP\cdot BCR}{BCP+BCR}$$

Отметим, что BCP (BCubed Precision) и BCR (BCubed Recall) не удовлетворяют сразу четырем принципам метрик, а F-мера — удовлетворяет. Таким образом далее будем оценивать модель с помощью метрик, описанных выше, опираясь при этом больше именно на F-меру.

3 Вычислительный эксперимент

3.1 Данные

Имеется набор документов. В каждом документе присутствует некоторое число новостных текстов. Каждый новостной текст размечен на некоторое число кластеров - мнений (данное число для каждого документа известно) с помощью инструмента Яндекс.Толока. Таким образом было получена выборка из 30 документов, на которых далее будут проводиться эксперименты.

Примеры рассматриваемых тем:

- обеспокоенность генсека НАТО из-за сотрудничества Москвы и Минска,
- в екатеринбурге пешеход выстрелил в лицо водителю, который ехал по тротуару,
- сообщения о росте заболеваемости covid-19 в Петербурге и др.

Примеры поляризации, в новостях о том, что генсек НАТО обеспокоен из-за сотрудничества Москвы и Минска:

- «Угроза от Минска и Москвы»? Путин заочно «успокоил» Столтенберга. Президент России Владимир Путин уже "успокоил"главу НАТО Йенса Столтенберга. Ещё до высказывания главы альянса об угрозе, исходящей от Москвы и Минска, Путин заявил, что руководство НАТО "в опасности раз позволяет себе такие заявления. ... Также в НАТО испугались дружбы России и Белоруссии. Как заявил генеральный секретарь альянса Йенс Столтенберг, там обеспокоены растущим сотрудничеством двух наших стран. В НАТО заявили, что готовы оборонять своих союзников от угроз, которые исходят от России или Белоруссии. ...
- Столтенберг заявил об обеспокоенности сотрудничеством России и Белоруссии. Генеральный секретарь НАТО Йенс Столтенберг заявил, что в альянсе обеспокоены сотрудничеством между Россией и Белоруссией. © Reuters «Мы бдительны и очень внимательно следим за тем, что происходит в Белоруссии», — сказал Столтенберг в интервью изданию Welt am Sonntag.По его словам, в НАТО готовы при необходимости защищать союзников от «любой угрозы, исходящей от Минска и Москвы».Он отметил, что альянс обеспокоен более тесным сотрудничеством России и Белоруссии в последние месяцы.Ранее Йенс Столтенберг заявил, что лидеры НАТО на саммите 14 июня в Брюсселе планируют обсудить действия России у границ Украины. ...

В данном примере первая новость относится к мнению России в данном вопросе, а вторая к мнению НАТО о сотрудничестве Москвы и Минска и предпринимаемых действиях.

Используемые инструменты разработки. Модель реализована на языке python. Из данных были извлечены признаки с использованием библиотеки Natasha. Natasha позволяет быстро и качественно получить синтаксический и морфологический разбор текстов для русского языка. NER (Named Entity Recognition, распознавание именованных сущностей) также было получено с использованием этой библиотеки. Тематическая модель строилась с использованием библиотеки BigArtm. Также использовалась бибилиотека skitit-learn для построения кластеризации k-means (k-средних).

Предобработка текста. Для модели кластеризации К-средних проводится следующая преобработка:

- удаление нерелевантных символов (любые сиволы, которые не относятся к буквам),
- приведение всех слов к нижнему регистру,
- приведение к нормальной форме (лемматизация),
- удаление стоп-слов,
- TF-IDF преобразование (TF — term frequency, IDF — inverse document frequency).

TF-IDF преобразование основано на следующих формулах:

$$tf(w,d) = \frac{n_w}{\sum_k n_k},$$

где n_w - частота слова w в документе d.

$$idf(w,D) = \log \frac{|D|}{|\{\, d_i \in D \mid w \in d_i \,\}|},$$

где |D| - число документов, $|\{\, d_i \in D \mid w \in d_i \,\}|$ - число документов, в которых встречается слово w (когда $n_w \neq 0$)

Таким образом,

$$\text{tfidf}(t,d,D) = tf(t,d) \times idf(t,D)$$

Большой вес в данной модели получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употребления в других документах. То есть слова, которые содержатся в большом количестве документов, не будут играть существенную роль.

Для тематических моделей предварительная обработка текста не требуется, так как сначала выделяются признаки на основе синтаксического разбора всех предложений. После выделения всех признаков уже они приводятся к нормальной форме и используются в дальнейшем.

3.2 Подбор параметров

Для решения данной задачи необходимо найти и исследовать зависимость результата от следующих факторов:

- веса каждой из модальностей,
- коэффициенты регуляризации

Имеющаяся выборка документов не позволяет сделать честный подбор параметров с помощью, например, кросс-валидации, так как размеченное число новостей, относящихся к одному политическому событию, не так велико. Поэтому было решено взять следующие веса модальностей:

Субъекты	Объекты	Положительные	Отрицательные	Именованные сущности
0.5	0.5	0.5	0.5	1

Таблица 1. Рассматриваемые веса модельностей

Вес регуляризации возьмем равным 2.

3.3 Результаты

В Таблице 2 представлены результаты работы алгоритма жесткой кластеризации K-средних. Анализируются модели, использующие следующие признаки:

- из новостного текста выделяются слова и биграммы, на которых строится модель TF-IDF, позволяющая перевести тексты в цифровые значения,
- из новостного текста выделяются признаки: субъекты/объекты, тональные слова и именованные сущности. Основываясь на этих словах строится модель TF-IDF.

	precision	recall	F1-мера
K-Means на полном тексте новости	0.686	0.665	0.666
K-Means на выделенных признаках	0.702	0.655	0.671

Таблица 2. Качество модели кластеризации K-средних на лексических и выделенных в данной работе признаках, преобразованных с помощью TF-IDF

Стоит заметить, что модель, основанная только лексических признаках (используются только слова без учета порядка) показывает качество немного хуже, чем модель, построенная на выделенных признаках. Это согласуется с предположением, что мнение является комбинацией некоторых более сложных сущностей, чем слова этого текста.

Далее рассмотрим тематическую модель, основанную на различных признаках. Результаты работы представлены в Таблице 3.

Заметим, что полученное качество тематической модели, основанной даже на одном из рассматриваемых признаков, сравнимо и немного выше качества модели K-средних. Таким образом, это обосновывает использование более сложного алгоритма классификации - тематической модели. Также отметим, что модель, основанная только

	precision	recall	F1-мера
Тональность(тональные слова)	0.602	0.828	0.68
Субъекты|объекты	0.594	0.817	0.669
Именованные сущности	0.651	0.725	0.675

Таблица 3. Значения метрик на различных выделенных признаках

на тональных словах показывает лучший результат. Это основано на том, что авторы новостей очень часто используют тонально окрашенные слова, описывая то или иное событие. И после анализа распределения таких слов можно сделать вывод о том, что тексты относятся к разным мнениям.

В Таблице 4 представлены различные комбинации признаков для тематической модели. Ожидается, что модель, основанная на большем числе признаков будет иметь лучший результат.

	precision	recall	F1-мера
Тональность + Субъекты|объекты	0.608	0.783	0.663
Тональность + Именованные сущности	0.617	0.791	0.674
Субъекты|объекты + Именованные сущности	0.626	0.746	0.672
Все признаки вместе	0.63	0.805	0.698

Таблица 4. Различные комбинации признаков и качество на них

Заметим, что качество модели, основанной на всех признаках сразу заметно выше, чем на моделях, основанных на каком-то подмножестве признаков. Это говорит о том, что признаки, рассматриваемые в отдельности не имеют возможности точно определить мнение, они выделяют какую-то часть выраженного мнения. А их комбинация использует максимальное количество извлеченной информации.

Заметим, что до этого рассматривались все сразу документы без учета количества мнений в них. А в имеющейся выборке число мнений может быть различным. Структура выборки при разбиении по числу мнений представлена на на Рис. 4

Объединим имеющуюся выборку документов по числу выделенных мнений в них. Таким образом получим три подвыборки:

- имеется 2 различных мнения в документе: 16 документов,
- имеется 3 различных мнения в документе: 9 документов,
- число мнений в документах больше трех: 5 документов.

Качество модели на данных подвыборках представлено в Таблице 5.

Таким образом, удалось построить модель, которая выполняет:

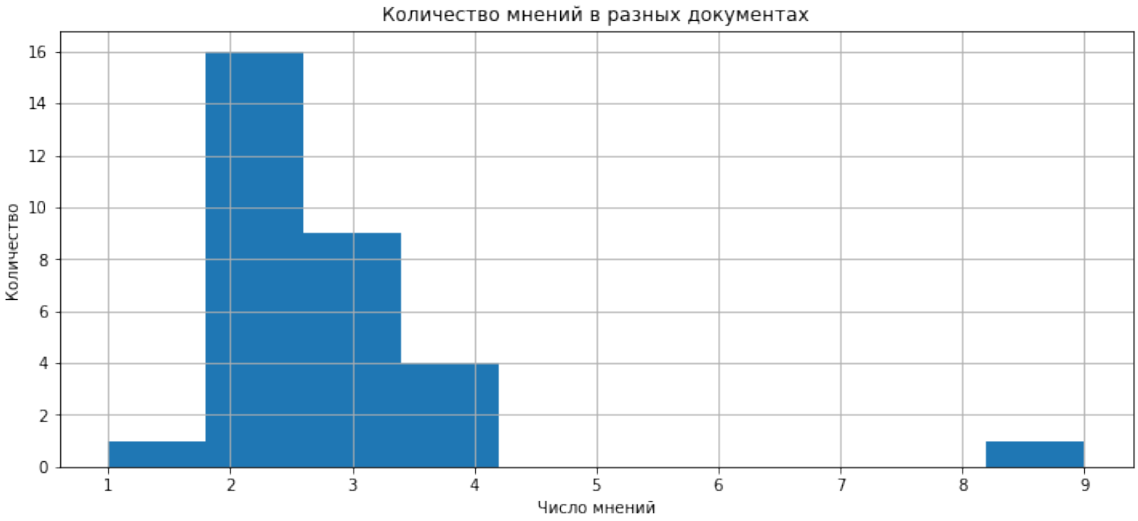

Рис. 4. Распределение числа мнений по выборке документов

	precision	recall	F1-мера
2 мнения	0.676	0.868	0.753
3 мнения	0.561	0.658	0.6
$>$ 3 мнений	0.5321	0.828	0.64

Таблица 5. Значения метрик на подвыборках, полученных на основе разбиения имеющейся выборки по числу мнений

- выделение необходимых признаков (субъекты, объекты, положительно и отрицательно окрашенные слова, именованные сущности) из новостных текстов,
- обучение тематической модели,
- кластеризацию с использованием данных, полученных с помощью тематического моделирования.

4 Заключение

В данной работе удалось построить модель обучения без учителя, решающую задачу поиска поляризации мнений в новостных текстах. Данный метод выделяет из текста следующие признаки: субъекты и объекты, тонально окрашенные слова и словосочетания, а также именованные сущности. Затем с помощью вероятностной тематической модели производится кластеризация текстов на несколько мнений. Качество модели, построенной на семантических и синтаксических, превосходит модель, использующую только лексические признаки. При этом модель, основанная на всех признаках сразу показала лучшие результаты.

5 Список используемой литературы

Список литературы

[1] Feldman D. G., Sadekova T. R., Vorontsov K. V. COMBINING FACTS, SEMANTIC ROLES AND SENTIMENT LEXICON IN A GENERATIVE MODEL FOR OPINION MINING, 2020

[2] Воронцов К. В «Вероятностное тематическое моделирование», 2013

[3] Kulagin D.I. «Publicly available sentiment dictionary for the Russian language»

[4] Большакова Е.И., Воронцов К.В., Ефремова Н.Э.Б Клышинский Э.С., Лукашевич Н.В., Сапин А.С Автоматическая обработка текстов на естественном языке и анализ данных, НИУ ВШЭ, 2017

[5] Amigó, Enrique, et al.: A comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints. In: Information Retrieval 12.4 (2009): 461-486.