ФЕДЕРАЛЬНОЕ АГЕНТСТВО ВОЗДУШНОГО ТРАНСПОРТА

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
ГРАЖДАНСКОЙ АВИАЦИИ

КАФЕДРА ПРИКЛАДНОЙ МАТЕМАТИКИ

**Курсовая работа**

**«Анализ тональности текста в обзорах фильмов»**

по дисциплине **«Математические модели»**

Выполнила:

студентка группы ПМ3-1

Руденко В. Ю.

Проверил:

заведующий кафедрой,

д.т.н., профессор

Кузнецов В.Л.

**Содержание**

1. [Введение 3](#_Toc8365778)

* [Обзор предметной области 3](#_Toc8365779)
* [Подходы к классификации 3](#_Toc8365780)
* [Практическое применение 5](#_Toc8365781)

1. [Постановка задачи 6](#_Toc8365782)

* [Описание данных 6](#_Toc8365783)
* [Выбор метрики качества 6](#_Toc8365784)

1. [Построение математической модели 6](#_Toc8365785)

* [Векторная модель текста 6](#_Toc8365786)
* [Приближения модели 7](#_Toc8365787)
* [Bag of Words + TF-IDF 7](#_Toc8365788)
* [Word2Vec 8](#_Toc8365789)

1. [Описание метода решения задачи 8](#_Toc8365790)
2. [Сравнение описанных моделей 9](#_Toc8365791)
3. [Заключение 10](#_Toc8365792)
4. [Литература 11](#_Toc8365793)

# Введение

### Обзор предметной области

Обработка естественного языка (NLP) — общее направление искусственного интеллекта и математической лингвистики. NLP изучает проблемы компьютерного анализа и синтеза естественных языков и представляет собой огромный спектр задач разного уровня. Одной из главных задач NLP является анализ тональности текста.

Анализ тональности текста – это область компьютерной лингвистики, занимающаяся выделением из текстов эмоционально окрашенной лексики или эмоциональной оценки автора.

Основной целью анализа тональности является нахождение мнений в тексте и выявление их свойств. Какие именно свойства будут исследоваться, зависит от поставленной задачи. К примеру, целью анализа может быть автор, то есть лицо, которому принадлежит мнение.

Основной же задачей анализа тональности текста является его классификация по тональной оценке. Под тональной оценкой или тональностью понимается эмоционально окрашенная лексика и эмоциональная позиция, выраженная автором относительно чего-либо. Тональность может быть позитивной, негативной или нейтральной. Тональность всего текста в целом можно определить как функцию (в простейшем случае сумму) лексических тональностей составляющих его единиц (предложений, слов, словосочетаний) и правил их сочетания. В основном используется бинарная классификация текста, то есть применяется два типа тональных оценок: позитивная и негативная.

### Подходы к классификации

Существует несколько подходов к классификации тональности [1]:

1. Подход, основанный на правилах.

Цель данного подхода — поверхностный синтаксический анализ на основе какого-либо правила, которое будет определять тональность текста. Для этого текст разбивается на слова или последовательности слов. Затем полученные данные используются для выделения часто использующихся шаблонов, которым присваивается позитивная или негативная оценка. Такой подход показывает хорошую точность при большом количестве правил.

Пример такого подхода: если высказывание содержит прилагательные из набора ['хороший', 'добрый', 'новый'] и не содержит прилагательных из набора ['плохой', 'старый', 'ужасный'], то высказывание является положительным.

1. Подход, основанный на заранее составленных тональных словарях с применением лингвистического анализа [2].

Словарь представляет собой список слов с приписанной им тональностью. Для его составления сначала отрабатывает отдельный лингвистический модуль, автоматически производящий морфологический анализ текста. Затем все слова размечаются по заранее подготовленным спискам тональной лексики. Каждому слову приписывается два атрибута, указывающие на тональность и/или силу тональности. Если слово не нашлось в списках тональной лексики, то оно считается нейтральным. После этого запускается первичный синтаксический анализ: слова и словосочетания объединяются в тональные цепочки, в предложении выделяются субъект, предикат и объект. Затем подсчитывается общая тональность текста.

1. Машинное обучение без учителя.

TF или частота слова - это отношение количества вхождения конкретного термина к суммарному набору слов в исследуемом тексте (документе). Этот показать отражает важность (весомость) слова в рамках определенной статьи/публикации.

IDF или обратная (инвертированная) частота документа - это инверсия частотности, с которой определенное слово фигурирует в коллекции текстов (документов). Благодаря данному показателю можно снизить весомость наиболее широко используемых слов (предлогов, союзов, общих терминов и понятий).

Приведем формулу для расчета обратной частоты документа:

,

где D — количество всех документов, d — количество документов, в которых содержится данное слово. Таким образом, чем чаще слово встречается, тем меньше его IDF.

На отношении этих показателей (TF/IDF) и основан данный подход, то есть наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте, и в то же время присутствуют в небольшом количестве текстов всей коллекции (показатель TF/IDF). Выделив эти термины и определив их тональность, можно сделать вывод о тональности всего текста целиком.

1. Машинное обучение с учителем [3].

Суть данного подхода заключается в построении статистического или вероятностного машинного классификатора (например, байесовского) на заранее размеченных текстах, а затем в его использовании при анализе новых текстов.

Краткий алгоритмданного подхода:

* собирается коллекция текстов, на основе которых обучается машинный классификатор;
* каждый текст раскладывается в виде вектора признаков, по которым в дальнейшем он будет исследоваться;
* указывается правильный тип тональности для каждого текста;
* производится выбор алгоритма классификации и метод для обучения классификатора;
* полученная модель используется для определения тональности текстов новой коллекции.

1. Комбинации перечисленных подходов.

Данный подход сочетает все или несколько подходов, рассмотренных выше, и заключается в применении классификаторов на их основе в определенной последовательности.

Примеры таких комбинаций:

* TF-IDF + logistic regression;

В работе [12] набор текстовых данных представляют в векторном виде, используя модель «мешка слов», так как она является простейшей. Затем для нормализации к полученным векторам применяется преобразование TF-IDF и проводится классификация текстов с помощью метода логистической регрессии на уже преобразованных данных.

* Bag of Words + naive bayes;

В работе [9] авторы рассматривают каждое текстовое сообщение в виде набора слов (Bag of Words). Для решения задачи классификации используются метод на основе наивного байесовского классификатора. Основные преимущества – простота реализации и низкие затраты при обучении и классификации.

* Delta TF-IDF + SVM;

В работе [11] рассматривается векторное представление текста с помощью «мешка слов» и с использованием функции взвешивания Delta TF-IDF. В качестве классификатора использовался метод опорных векторов (SVM) с линейным ядром, потому что оно дает более высокую точность.

* Word2Vec + Random Forest.

В работе [10] для векторного представления русскоязычного текста используют технологию Word2Vec. В качестве классификатора используется Random Forest - случайный лес с количеством деревьев равным 100.

Word2Vec –технология от Google, использующаяся для статистического анализа больших массивов текстовой информации. Она собирает статистику по совместному появлению слов в фразах, после этого с помощью нейронных сетей решает задачу уменьшения размерности и в итоге выдает компактные векторные представления слов, достаточно полно отражающие отношения этих слов в обрабатываемых текстах [13].

### Практическое применение

Анализ тональности текста имеет важное практическое применение и находит его во множестве областей. К примеру, в бизнес сегменте, социальных и политических исследованиях:

* определения уровня лояльности потребителя к продукту. На основе данных мониторинга социальных систем делаются выводы о популярности того или иного продукта, нахождение текущих трендов среди покупателей;
* анализ данных о политических позициях пользователей, прогнозирование результатов выборов;
* на основе анализа тональности текстов новостных лент, обзоров финансовых аналитиков, отчетов трейдеров, а также общего настроения пользователей социальных сетей определяется корреляция этих данных с трендами фондовых рынков, и строятся прогнозы изменения цен финансовых активов.

Так же анализ тональности тестов используется для борьбы с киберпреступностью. Например, используется автоматический способ идентификации людей, склонных к педофилии, на основе анализа текстовых сообщений в социальных сетях [8].

Приведем примеры готовых программных продуктов, использующих анализ тональности текстов:

* Tweet Sentiment Visualization App [5];

Этот продукт позволяет анализировать информацию о продукте, который упоминают пользователи, при помощи данных из веб-сервиса Twitter. Пользователь получает в ответ на свой запрос подборку позитивных, негативных или нейтральных микросообщений. Сервис визуализирует соответствующий результат при помощи инфографики. Для анализа тональности данный веб-сервис использует тональный словарь.

* Text Analytics API [6].

Text Analytics API представляет собой набор веб-служб для анализа текста, созданных с использованием алгоритмов машинного обучения Microsoft. API можно использовать для таких задач, как анализ настроений, извлечение ключевых фраз и определение языка.

# Постановка задачи

В качестве предметной области, представляющей входные данные, была выбрана тема обзоров фильмов. Необходимо по заданному отзыву, определить к какому классу относится фильм, к классу «хороший» или «плохой».

### Описание данных

Входные данные представляют собой файл:

* labeleledTrainData.tsv – маркированный обучающий набор. Файл разделен табуляцией, имеет строку заголовка, за которой следует 25 000 строк, содержащих идентификатор, тональность и текст для каждого отзыва.

### Выбор метрики качества

Так мы рассматриваем задачу бинарной классификации и наши данные являются сбалансированными, то для оценки качества работы алгоритма выберем площадь под ROC-кривой (AUC-ROC). Данный показатель часто используется для сравнительного анализа нескольких моделей классификации.

ROC-кривая – графическая характеристика качества бинарного классификатора, которая характеризует зависимость доли верных положительных классификаций (TPR) от доли ложных положительных классификаций (FPR).

Доли TPR и FPR можно вычислить по формулам:

Площадь под ROC-кривой является объединенной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Численный показатель площади под кривой называется AUC. Вычислить его можно с помощью численного метода трапеций:

Чем больше показатель AUC, тем выше предсказательная способность модели [7].

# Построение математической модели

После того, как мы провели предварительную обработку текстов отзывов необходимо преобразовать их в какое-либо числовое представление для дальнейшей работы. Для решения поставленной задачи будем использовать две векторные модели: Bag of Words с преобразованием TF-IDF и Word2Vec.

### Векторная модель текста

Векторная модель (vector space model) — представление текстов векторами из одного общего для всей коллекции текстов векторного пространства.

Векторная модель является основой решения задач информационного поиска: поиск документа по запросу, классификация документов, кластеризация документов и т.д. Текст в векторной модели рассматривается как неупорядоченное множество слов.

Вектор, являющийся модельным представлением текста в векторном пространстве, образуется упорядочением весов всех слов (включая те, которых нет в конкретном тексте). Размерность этого вектора, как и размерность пространства, равна количеству различных слов во всей коллекции, и является одинаковой для всех текстов коллекции.

Формально:

,

где — векторное представление jо текста, — вес -го слова в -м документе, — общее количество различных слов во всех документах коллекции.

Для полного определения векторной модели необходимо указать, каким именно образом будет отыскиваться вес слова в документе [16].

### Приближения модели

В качестве основных приближений моделей приведем описание предварительной обработки данных - процесса очистки и подготовки текстов к классификации.

Предварительная обработка текста включает в себя:

* удаление HTML тегов;
* удаление всех не буквенных символов (знаков препинания, чисел и т.д.) и замена их на пробелы;
* конвертация слов в нижний регистр;
* удаление стоп-слов, к примеру, таких как «a», «and», «is» и т.д. Так же удалим слова «movie» и «film», так как, скорее всего, они будут довольно часто встречаться в отзывах, но они явно не несут эмоциональной окраски.

Кроме всего выше перечисленного к приближениям исследуемых моделей так же относится и то, что все отзывы написаны на английском языке и объединены единой тематикой.

### Bag of Words + TF-IDF

Модель Bag of Words («мешок слов») – самая простая модель представления текста как набора слов без учета их взаимного расположения и взаимных связей. Модель сводит текст к вектору, где каждая его позиция представляет слово, а значение этой позиции представляет число раз, которое это слово используется в тексте. После преобразования текста в «мешок слов» мы можем рассчитать различные меры для характеристики текста [15].

Модель «мешка слов»широко используется в методах классификации документов, где частота появления каждого слова в тексте используется в качестве признака для обучения классификатора. Однако частоты терминов не являются лучшим представлением для текста. Иногда слова могут встречаться во многих документах текстовой коллекции. Следовательно, они не могут характеризовать принадлежность документа тому или иному классу, так как не являются ключевыми.

Добавим дополнительное преобразование TF-IDF. Данное преобразование используется для корректировки значений вектора в соответствии с числом рецензий, использующих слово. Слова, встречающиеся во многих отзывах, могут быть менее значимыми, чем слова, встречающиеся реже. Преобразование TF-IDF уменьшает значение данного слова пропорционально количеству документов, в которых оно появляется.

Важный недостаток такой модели в том, что тексты как наборы слов проецируются в пространство высокой размерности, что обусловлено объемом используемого словаря. Данный недостаток характеризуется экспоненциальным ростом сложности вычислений из-за увеличения размерности данных.

В используемых данных представлено большое количество отзывов, что даст нам составить довольно объемный словарь даже с учетом удаления стоп-слов. Для ограничения размера векторов отзывов выберем максимальный размер словаря. Ограничимся 5000 наиболее часто встречающихся слов.

### Word2Vec

Как уже отмечалось выше Word2Vec – это модель, которая с помощью нейронной сети выдает компактные векторные предоставления слов. Векторное представление слова основывается на контекстной близости: слова, имеющие схожий смысл в векторном представлении будут иметь близкие координаты векторов-слов.

Алгоритм работы модели [14]:

* читается корпус, и рассчитывается встречаемость каждого слова в корпусе;
* массив слов сортируется по частоте и удаляются редкие слова;
* строится дерево Хаффмана для кодирования словаря;
* из корпуса читается субпредложение - базовый элемент корпуса и проводится процесс изъятия наиболее частотных слов из анализа;
* по субпредложению проходим окном - максимальной дистанцией между текущим и предсказываемым словом в предложении;
* применяется нейросеть.

Существенное отличие данной модели от модели Bag of Words в том, что она так же решает задачу уменьшения размерности. Чтобы добиться наиболее эффективной работы Word2Vec, необходимо использовать большие текстовые корпусы для его обучения. Это позволяет повысить качество предсказаний. Поэтому предобработка данных для Word2Vec немного отличается от предыдущей модели, так как нет необходимости удалять стоп-слова.

# Описание метода решения задачи

В качестве метода решения поставленной задачи выберем логистическую регрессию, так как она хорошо подходит для решения задач бинарной классификации.

Логистическая регрессия – это статистический метод, используемый для предсказания вероятности события по значениям множества признаков. Для этого применяется логистическая функция. Практическое значение логистической регрессии заключается в том, что она является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных. Иными словами, с помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого.

Для этого вводится так называемая зависимая переменная, принимающая два значения: 0 - текст окрашен негативно или 1 - позитивно. Также вводится множество независимых переменных - отдельные слова из отзыва, на основе которых вычисляется вероятность принятия того или иного значения зависимой переменной.

Делается предположение, что вероятность наступления события:

где – стандартное уравнение регрессии, имеющее вид .

Если > 0.5, то документ считается положительно окрашенным, иначе – негативно[7].

# Сравнение описанных моделей

Сравнение двух моделей будем проводить по оценке качества их работы. Для обеих моделей мы используем одни и те же входные данные, в которых находится 25000 отзывов. Для оценки качества разобьем исходные данные на обучающую и тестовую выборку. Так для обучения моделей будем использовать 18750 отзывов и 6250 для тестирования.

Посчитаем значения выбранной метрики качества на тестовых выборках двух моделей и построим ROC-кривые.

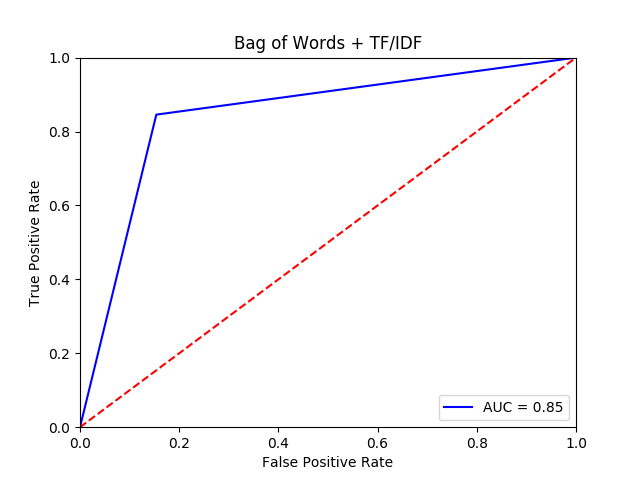


Рис.1 – Точность классификатора logistic regression по метрике AUC для модели Bag of Words + TF-IDF (при 5000 признаков).

Значение метрики AUC для модели Bag of Words + TF-IDF: .

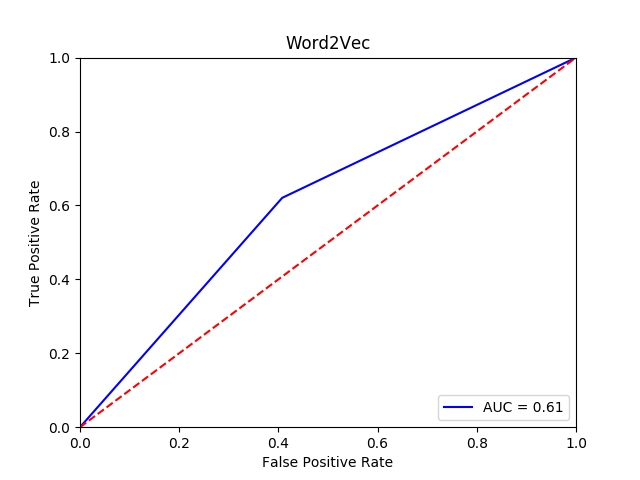


Рис.2 – Точность классификатора logistic regression по метрике AUC для модели Word2Vec.

Значение ROC-AUC для модели Word2Vec:

Анализируя полученные результаты можно заметить, что оценка качества более простой модели Bag of Words + TF-IDF несколько выше, чем у модели Word2Vec. Возможно, это связанно с тем, что использованное количество отзывов было недостаточным для второй модели, так как, как уже говорилось ранее, чтобы повысить качество ее работы необходимо использовать большие текстовые наборы данных для его обучения

# Заключение

В данной работе рассматривалась задача бинарной классификации тональности отзывов к фильмам.

Общий алгоритм тональности:

* предварительная обработка данных;
* построение вектора признаков;
* применение метода машинного обучения с учителем;
* оценка полученных результатов.

Были рассмотрены две модели векторного представления текста: Bag of Words с преобразованием TF-IDF и Word2Vec. В ходе решения поставленной задачи «мешок слов» показал несколько превосходящие результаты над моделью Word2Vec. Однако стоит отметить, что в модель Bag of Words использовался словарь с 5000 наиболее часто встречающимися словами.

# Литература

1. Юрганов Александр Александрович Сентимент-анализ как инструмент исследования текстов // Проблемы Науки. 2017. №29 (111). URL: https://cyberleninka.ru/article/n/sentiment-analiz-kak-instrument-issledovaniya-tekstov (дата обращения: 09.05.2019).
2. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке.
3. B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, 2002, pp. 79–86.
4. Sentiment Analysis. Nearly Everything You Need to Know // URL: https://monkeylearn.com/sentiment-analysis/ (дата обращения: 10.05.2019)
5. Tweet Sentiment Visualization App // URL: https://www.csc2.ncsu.edu/faculty/healey/tweet\_viz/tweet\_app/
6. Text Analytics API - // URL: https://westus.dev.cognitive.microsoft.com/docs/services/TextAnalytics.V2.0/
7. Паклин Н. Логистическая регрессия и ROCанализ – математический аппарат. – BaseGroup Labs, 1995 –2009.
8. Якушенкова А.Д. Применение анализа тональности текста для борьбы с киберпреступностью // Образование и наука в современных условиях : материалы VII Междунар. науч.–практ. конф.– Чебоксары: ЦНС «Интерактив плюс», 2016. – С. 97-99. – ISSN 2412-0537.
9. Юсупова Н.И., Богданова Д.Р., Бойко М.В. Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Вестник УГАТУ 2012. №6 (51).
10. Мишенин А.Н., Нефедова Е.А. Анализ тональности текстов с использованием технологии Word2Vec // Естественные и математические науки в современном мире: сб. ст. по матер. XLIV междунар. науч.-практ. конф. № 7(42). – Новосибирск: СибАК, 2016. – С. 89-97.
11. Martineau J. C., Finin T. Delta tfidf: An improved feature space for sentiment analysis //Third international AAAI conference on weblogs and social media. – 2009.
12. 85. Using free text for classification – ‘Bag of Words’ - // URL: https://pythonhealthcare.org/2018/06/02/85-using-free-text-for-classification-bag-of-words/ (дата обращения: 10.05.2019)
13. https://code.google.com/archive/p/word2vec/
14. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR, 2013
15. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). "Learning Word Vectors for Sentiment Analysis." *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*.
16. Моченов С. В., Бледнов А. М., Луговских Ю. А. Векторная модель представления текстовой информации //Материалы международной научной конференции. – 2006. – С. 133-139.