Курсовая работа на тему: "Анализ тональности текста методами машинного обучения"

Пехтерев Д.О.

Аннотация.
Было проведено исследование предыдущих работ для определения того, какие методы являются современными, какие их особенности уже были выявлены, а также для определения вектора развития данной области. В расчётной составляющей происходит изучение алгоритма по анализу тональности текстовых данных, где используется метод логистической регрессии. Для выделения признаков были рассмотрены три различных метода: Мешок слов (от англ. BoW), TF-IDF и N-граммы, в итоге же использовался только TF-IDF. Данные, используемые в работе были собраны пользователем Julian McAuley с сайта Amazon за период: с мая 1996 по июль 2014 года, где мною были выбраны различные категории товаров, такие как: Рецензии на книги, Здоровье и личная гигиена, Кино и Телевидение и Электроника.

Для цитирования: Пехтерев Д. О., "Курсовая работа", *НИС*, **0**:0 (2000), 1–0.

Об авторах:
Пехтерев Денис Олегович, студент направления: "Суперкомпьютерное моделирование в науке и инженерии"
НИУ ВШЭ МИЭМ им. А.Н.Тихонова
ул. Мясницкая, 20, г.Москва, Россия, 101000
e-mail: dopekhterev@edu.hse.ru

Введение

0.1. Актуальность работы

Мы живём в то время, когда интернет есть в каждом нашем устройстве, независимо от того, где мы находимся и чем занимаемся. Что касается России, то к началу 2019 года количество интернет-пользователей среди населения 16+ стало равно 90 миллионов человек, что есть 75,4% взрослого населения страны. [1] Проникая во всё более новые отрасли нашей жизни, интернет стал наполняться огромным количеством информации, генерируемой потребителями продуктов. Стоит заметить, что почти у каждого конкурентноспособного бренда есть как минимум свой сайт или форум, поэтому клиенты имеют возможность расписать свой личный опыт использования того или иного продукта. Если смотреть на этот процесс со стороны успешного функционирования бизнеса, то одна из его важнейших частей – понимание клиента и его потребностей. Приблизительно 95% потребителей в возрасте от 18 до 34 лет читают отзывы на продукт или услугу перед покупкой, из них 91% доверяют онлайн отзывам так же, как и личным рекомендациям. [2]

Обычно перед компанией стоит задача по мониторингу реакции клиента на:

- Выпуски новых продуктов (запуск новых сервисов);

- Внесение изменений в существующие продукты (услуги);
- Внедрённые или готовящихся к внедрению инициатив, связанных с обслуживанием клиентов;
- Запуск новых рекламных компаний;

Наиболее же важным для компаний является определение того, какие их действия вызывают положительный отклик у клиентов, а какие отрицательный.

Зачастую, количество отзывов по конкретным продуктам (услугам), может превышать 10000 записей в день. Данная цифра включает в себя также и сообщения, поступающие в службу поддержки, которые необходимо сразу же сортировать и классифицировать, ведь от этого зависит опыт использования продукта клиентом и его мнение о том, на сколько дорог он как клиент для компании. Одной из задач, которая стоит перед компаниями сейчас – является задача по определению тональности текста (в пер. с англ. sentiment analysis).

Сейчас решением данной задачи на Российском рынке занимаются два крупных сервиса: YouScan и Brand Analytics. С одной стороны, эти системы кажутся абсолютно схожими – они решают одинаковые задачи и дают примерно одинаковый анализ. Главной же проблемой данных решений является их цена и ограниченный функционал. Рассмотрим ценообразование, которое используется сервисами YouScan и Brand Analytics по минимальной стоимости:

- Standard , YouScan – 35 000 руб / мес
  - за каждые 5 тем в месяц
  - 100 000 упоминаний в каждой теме
- Стартовый плюс , Brand Analytics – 35 000 руб / мес
  - сообщений в аккаунте за месяц: 20 000 сообщений
  - количество тем: 5
  - автоматическая разметка тематик (тегов на тему): 30[1]

Как мы видим, сервисы по своей сути предлагают одинаковый пул услуг: ограниченный набор тем (вы можете отслеживать только 5 услуг/товаров/запросов) и имеют ограничение на количество упоминаний (другими словами, за эти деньги, больше указанного выше лимита вы увидеть не сможете). Большим и достаточно развитым бизнесам не составит труда купить подписку подороже и отслеживать свои товары на ежедневной основе, получая при этом достаточно хороший сервис. Однако здесь оказывается незанятой другая ниша рынка – компании, имеющие достаточно много продуктов (брендов 50), но не выделяющих большие бюджеты на анализ отзывов, оставляемых относительно их продуктов. Такой анализ был бы полезен продакт-менеджерам [2] для того, чтобы отслеживать реакцию потребителей и смотреть на то, как они воспринимают/потребляют продукт, и с какими сложностями они сталкиваются. К таким компаниям относятся:

[1] например, бывают тэги: головная боль, кожный зуд и т.д.

[2] Люди, возглавляющие все активности, связанные с продуктом

- Фармацевтические компании;
- Медицинские клиники;
- Агрегаторы такси;

0.2. Поставновка цели и задачи

Цель работы: применение современных средств прикладной математики для анализа данных

Задача работы: Анализ социальной активности потребителей в интернете, касающейся тональности оставленных ими отзывов

Основная задача: Анализ тональности отзывов методами машинного обучения

Подходы: Машинное обучение.

Используемые методы:

- Метод логистической регрессии
- Метод опорных векторов
- Метод гребневой регрессии

1. Обзор литературы

В этой части работы будут проанализированы статьи других авторов, напрямую связанные с анализом тональности текста, анализом текста, содержащегося в отзывах о покупках на различных сайтах и анализом тональности текста, содержащегося в социальных сетях, таких так Twitter.

В обзорной статье [3] авторы изучили основные алгоритмы как по выделению признаков, так и по построению моделей, способных определять тональность текста. Говоря об анализе тональности, авторы обращаются к рисунку 1, который описывает существующие и использующиеся сейчас техники по анализу тональности текста.

Также авторы рассмотрели основные методы по извлечению признаков из текста:

1. Методы, в основе которых лежит измерение корреляции между словами и классами:
    - Метод поточечной взаимной информации (от англ. Point-wise Mutual Information (PMI)): метод основан на введении и использовании меры, которая показывает контекстную связь между признаками и классами. Пример использования: метод использовался в работе Yu и Wu [4], где авторы при помощи данного метода расширяли первоначальный набор слов, который был собран из небольшого набора текстов, посвященных новостям с фондового рынка. Результаты их модели превзошли результаты других разновидностей моделей PMI;

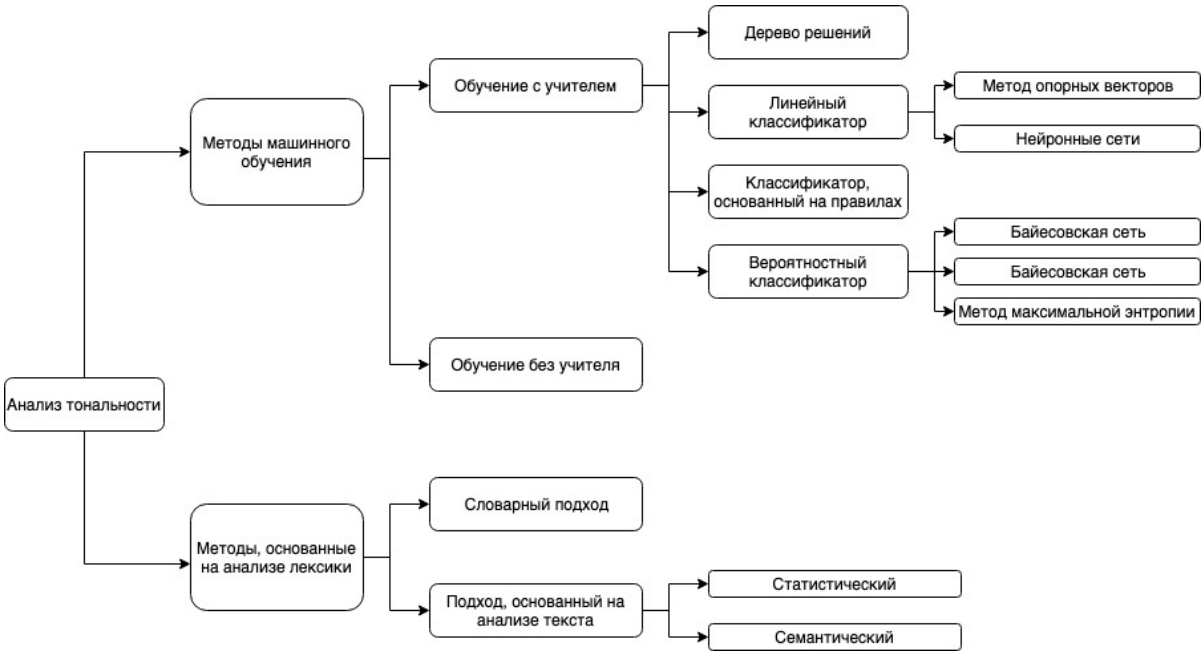

Рис. 1. Техники по анализу тональности текста (на основе оригинального рисунка 2 из статьи [3])

- Хи-квадрат (от англ. Chi-square (χ^2)): метод также отвечает за поиск величины связи конкретного слова и класса, однако лучше PMI в том, что внутри него используются нормализованные значения, соответственно, эти значения больше подходят к терминам в той же категории. Пример использования: в результате исследования Фаном и Чангом [5] контекстной рекламы, ими было показано, что их метод может выделить рекламные объявления, которые непосредственно коррелируют с интересами блогера. Для выделения признаков в работе использовался метод хи-квадрат, тогда как для модели классификации использовался метод опорных векторов, изображённый на рисунке 1;

2. Методы, суть которых заключается в уменьшении размерности данных:

   - Латентно-семантический анализ (от англ. Latent Semantic Indexing (LSI))

3. Методы, основанные на статистическом подходе:

   - Hidden Markov Model (HMM)
   - Latent Dirichlet Allocation (LDA)

4. Другие методы:

   - Метод Мешка слов (от англ. Bag of Words (BoW))

Авторы в [6] провели исследование по анализу тональности наборов данных, в состав которых входили: Cornell Movie Dataset, Twitter Dataset, Amazon Products Dataset и IMDB Dataset, используя различные методы машинного обучения. В работе рассматривались следующие методы: Naive Bayes (NB), стохастический градиентный спуск (Stochastic gradient descent - SGD), метод опорных векторов (support vector machine - SVM), пассивный агрессивный метод (Passive Aggressive - PA), максимальной энтропии (Maximum Entropy), Adaptive Boosting (AdaBoost), Multinomial Na?ve Bayes (MNB), Bernoulli Naive Bayes (BNB), Ridge Regression (RR) и Logistic Regression (LR). В результате использования данных методов вместе с методами по извлечению признаков из наборов данных: TF-IDF и N-Грамм (юниграмма, биграмма и триграмма), было получено, что использование алгоритмов PA и RR по сравнению с остальными методами дают одни из самых высоких результатов классификации: от 87% до 99.96%, полученный при использовании метода PA. Также было выделено, что метод логистической регрессии и метод опорных векторов также давали приемлемые результаты: 87.56% и 85.76% соответственно.

Авторы в [7] проводят анализ тональности текста (положительный/отрицательный), используя набор данных, состоящий из отзывов, оставленных пользователями на сайте Amazon об электрических товарах (Kindle, DVD и других) в период с февраля 2012 года по июль 2017 года. Для извлечения признаков из данных использовался метод TF-IDF, а для обучения модели были рассмотрены методы машинного обучения: 3 разновидности Naive Bayes классификатора: обычный метод Naive Bayes, Multinomial Naive Bayes и Bernoulli Naive Bayes, а также логистическая регрессия. В итоге были получены следующие результаты: Multinomial Naive Bayes - 92.87%, Bernoulli Naive Bayes - 92.35% и Логистическая регрессия - 93.34%.

Те же самые авторы в [8] изучали применение методов по анализу тональности текста, используя библиотеки языка программирования R по машинному обучению применительно к отзывам покупателей о приобретённой еде на сайте Amazon. В работе описан алгоритм при помощи которого проводилась классификация отзывов, а также приведены облаки слов, которые имели больший вес в модели при определении тональности.

Ещё 1 работа по анализу тональности отзывов с сайта Amazon является работа [9], где авторы предлагают алгоритм по обработке отзывов, призванный разбить оставленные пользователями отзывы на две категории:

- Отзывы о представленном сервисе;
- Отзывы о характеристиках продукта;

Далее авторами используется собственный алгоритм, который проводит анализ отзывов на предмет того, является ли отзыв прямым или косвенным. Например: “Камера не лучше, чем у iPhone 4s” - косвенный отзыв, фразы из предложений сравниваются с заранее выделенными в ходе ручной работы с отзывами фразами, по итогу чего и происходит классификация каждого отзыва. Также интересна обработка отрицаний в статье авторов, когда алгоритм встречает в отзыве фразу “не хороший товар”, то сначала алгоритм анализирует тональность слова хороший (стоит также отметить, что он ищет следующую связку: частицу, обозначающую отрицание и прилагательное). После чего, сначала, при помощи словаря тональности

определяется тональность прилагательного и, так как ещё есть частица, означающая отрицание, то данной фразе проставляется противоположная тональность.

Алгоритм работы с отзывами, после извлечения набора фраза из предложений, может быть представлен следующим образом:

- Все фразы проверяются на то, являются ли они косвенными или прямыми относительно продукта;
- Проверяются на наличие с частицей “не”;

2. Отчёт об использованных методах

2.1. Описание наборов данных

В данной работе использовался следующий набор данных: отзывы на категории товаров, представленные в разных категориях на сайте Amazon в период с мая 1996 по июль 2014 года.Набор данных был собран в файлы по категориям профессором Стэнфордского университета Джулианом Макаули. [12] Были взяты наборы данных по следующим категориям:

1. Отзывы на товары категории "Рецензии на книги"

   Всего отзывов в категории: 8 898 041 записей

2. Отзывы на товары категории "Электроника"

   Всего отзывов в категории: 1 689 188

3. Отзывы на товары категории "Кино и Телевидение"

   Всего отзывов в категории: 1 697 533

4. Отзывы на товары категории "Компакт-диски и винилы"

   Всего отзывов в категории: 1 097 592

5. Отзывы на товары категории "Kindle Store"

   Всего отзывов в категории: 982 619

Наборы данных были представлены в одинаковом формате JSON[3] и состояли из следующих колонок:

- reviewerID – уникальный идентификатор пользователя
- asin – уникальный идентификатор продукта
- reviewerName – имя человека, оставившего отзыв
- helpful – оценка полезности отзыва
- reviewText – текст отзыва

[3]JavaScript Object Notation - текстовый формат обмена данными, основанный на JavaScript.

- overall – рейтинг, оставленный пользователем (от 1 до 5)
- summary – краткий итог отзыва
- unixReviewTime/reviewTime – время, когда отзыв был оставлен пользователем

Из этих столбцов для нашей задачи будут интересны только reviewText - сам отзыв, оставленный посетителем сайта и overall - оценка пользователя, которая ставится в соответствии с его отзывом.

Итак, обработав отзывы в формате json и переведя их в формат pandas[4] - DataFrame, мы получили 5 наборов данных. Для начала будем рассматривать в каждом из этих наборов данных максимальное количество записей - 1 млн.

2.2. Обработка данных

Первым шагом в данном упражнении, переведём оценки из формата 1-5 в 2 класса, которые понадобятся нам для классификации (-1/1), для этого воспользуемся следующей логикой:

- при overall $\in$ (1,2) присвоим значение -1, что будет означать, что отзывы данной категории имеют отрицательную тональность;
- при overall $\in$ (3) не будем присваивать никакого значения, так как мы будем рассматривать задачу бинарной классификации;
- при overall $\in$ (4,5) присвоим значение 1, так как оставляя отзывы с такими оценками, мы можем быть уверены, что пользователь был удовлетворён своей покупкой, а значит отзыв имеет положительную тональность;

Имеем 5 наборов данных, состоящих из отзывов пользователей и их оценок, где в каждом наборе данных содержится приблизительно миллион записей. Так как в задачах машинного обучения важно иметь сбалансированные выборки, а в разных категориях товаров разное количество положительных и отрицательных отзывов, то была написана функция, которая считает в каждом наборе данных количество положительных и отрицательных отзывов, берёт наименьшее из них и обрезает класс, где отзывов больше, тем самым получается равное количество положительных и отрицательных отзывов в выборке.

Количество отзывов по классам тональности

Название категории	Положительные	Нейтральные	Отрицательные
"Рецензии на книги"	7 203 909	955 189	738 943
"Кино и Телевидение"	1 289 602	201 302	206 629
"Электроника"	1 356 067	142 257	190 864
"Компакт-диски и винилы"	903 002	101 824	92 766
"Kindle Store"	829 277	96 194	57 148

[4]Pandas - библиотека языка Python для работы с наборами данных

2.2.1. Векторизация отзывов (метод «Мешка слов»)

Первым этапом при анализе тех или иных текстов методами машинного обучения, текстовые данные требуется перевести в числовой формат.

Для этого существует определённых набор методик, таких как:

- Метод "Мешок слов"(BoW, bag-of-words) [10] - метод переводит набор данных в матричный формат, где каждая строка является отдельным предложением или отдельным отзывом, тогда как столбец отвечает за 1 слово в этом предложении/отзыве. Элементами же матрицы являются частоты с которыми данные слова встречаются в предложении.

- Метод TF-IDF (TF — term frequency, IDF — inverse document frequency) - метод, суть которого состоит в оценке важности слова как в контексте рассматриваемого экземпляра документа, так и относительно других документов, находящихся в коллекции.

  Частота слов (TF) вычисляется как отношение количества вхождений конкретного слова в документе к общему количеству слов в этом документе и записывается в следующем виде:

  $tf(a, d) = n_a / \sum_k (n_k)$,

  где n_a - количество вхождений слова a в документ n, $\sum_k (n_k)$ - общее число слов в документе n.

  Обратная частота документа (IDF) - отношение общего количества документов в коллекции $|D|$ к числу документов, где присутствует данное слово $|d_i \in D | a \in d_i|$ и записывается в виде:

  $idf(a, D) = log(|D|/|d_i \in D | a \in d_i|)$

  В конечном итоге, меру на которую данный метод опирается, можно записать следующей формулой: $tf - idf(a, d, D) = tf(a, d) * idf(a, D)$

- N-граммы (N-grams) - метод, основанный на методе Мешка слов, однако отличается тем, что данный метод позволяет учесть порядок слов в предложении, что является полезным свойством, особенно при анализе таких предложений как: "Продукт плох, совсем не качественный"или "Продукт качественный, совсем не плохой" [11]. Данные предложения будут иметь одинаковое числовое представление при их обработке методом Мешок слов. Для такого рода ситуаций подходит метод n-граммы, так как позволяет захватить контекст использования данного слова, учитывая не только количество отдельных слов (токенов), но и пар или словосочетаний состоящих из 3-ёх слов.

Удалив из данных знаки препинания и другие лишние символы (отмечу, что они могут быть исследованы при более глубоком анализе) и приведя все слова к нижнему регистру, я последовательно, используя 3 вышеописанных метода, создавал численные представления всех имеющихся отзывов.

2.2.2. Формальная постановка задачи обучения линейного классификатора

Формальная постановка рассматриваемой задачи: X - множество объектов, которые представляют из себя текстовые отзывы покупателей о приобретённом товаре. $x_1, ..., x_l \subset X$ - обучающая выборка. Рассматривается двухклассовая задача, $y_j \in (-1, 1)$ - признак, который отвечает за тональность отзыва: положительный/отрицательный. Имеем обучающую выборку: "отзыв - тональность которая может быть представлена в виде: $X^m = (x_1, y_1), ..., (x_m, y_m)$ В логистической регрессии строится линейный классификатор a типа [13]:

$a(x, w) = sign(\sum_{j=1}^{n} w_j * f_j(x) - w_0) = sign\langle x, w\rangle$,

где

- w_j - вес j-го признака
- w_0 - порог принятия решения
- $w = (w_0, w_1, ..., w_n)$ - вектор весов
- $\langle x, w\rangle$ - скалярное произведение признаков объекта на вектор весов

Задачей является настройка вектора весов w по заданной выборке X^m

Логистическая регрессия отличается функцией потерь, которая служит для решения этой задачи и отвечает за минимизацию эмпирического риска, имея вид:

$Q(w) = \sum_{i=1} ln(1 + exp(-y_i * \langle x_i, w\rangle)) \to \min\limits_{w}$

После того, как вектор w найден, можно вычислить как классификацию для объекта x, так и оценить апостериорные вероятности принадлежности объекта определённому классу:

$P(y|x) = \sigma(y\langle x, w\rangle), y \in Y$, где:

- $\sigma(z) = 1/(1 + e^{(} - x)$ - сигмоидная функция

2.2.3. Логистическая регрессия

Для решения данной задачи бинарной классификации, будем использовать логистическую регрессию, которая является одним из методов построения линейного классификатора и даёт оценивать апостериорные вероятности принадлежности объектов классам.

Так как логистическая регрессия является частным случаем линейного классификатора, то стоит упомянуть, что главная идея линейного классификатора заключается в том, что признаковое пространство может быть разделено гиперплоскостью на два отдельных полупространства. В каждом из полупространств прогнозируется одно (в случае бинарной классификации, из двух) значение целевого класса. Когда это можно сделать без ошибок, то выборка признаётся линейно разделимой.

Для решения данной задачи я использовал пакет sklearn, в котором содержится функция LogisticRegression. Мною использовалась штрафная функция типа $L2$, при том, что в пакете существует три варианта регуляризации:

- L1 регуляризация:
  $\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$.

- L2 регуляризация:
  $\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$.
- Elastic-Net регуляризация:
  $\min_{w,c} \frac{1-\rho}{2} w^T w + \rho \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$, где:
  - коэффициент p отвечает за силу l_1 регуляризации и l_2 регуляризации.

—

Также будет добавлено про то, как выбирался (подбирался) показатель регуляризации для классификации.

2.2.4. Оценка точности классификации

Исследователями выделяется большое количество оценки точности классификации.
К ним относятся:

- ACC ()
- Precision
- Recall
- F1-score
- ROC - кривая (кривая ошибок).
- Матрица ошибок (confusion matrix, матрица неточностей) -

Самым популярным методом является метод Accuracy: суть заключается в том, что при исследовании точности модели мы смотрим на долю правильных ответов, которая может быть записана в следующем виде: $Base\text{Rate} = \text{argmax} \frac{1}{l} \sum_{i=1}^{\text{l}} [y_o = y_i]$ Данную формулу также принято записывать в следующем виде: $Accuracy\text{Rate} = \frac{\text{TP+TN}}{TP+FP+FN+TN}$
В рассматриваемом примере, данный показатель будет вычисляться на основе сравнения результатов модели с заданными результатами последних 10000 отзывов в каждом наборе данных.

2.2.5. Результаты

Результаты работы классификатора:

Название категории	AccuracyRate, %
"Рецензии на книги"	91.47%
"Здоровье и личная гигиена"	85.78%
"Кино и Телевидение"	90.65%
"Электроника"	89.84%

Здесь мы можем увидеть результаты работы классификатора, реализованного с использованием логистической регрессии. Интересно изучить слова, которые оказали наибольшее влияние на работу классификатора. Так, например, при анализе

отзывов на категорию книги одним из важных факторов для модели при определении тональности стало наличие в предложении автора: Joseph Klausner. Это связано с тем, что под произведениями данного автора, представленными на сайта Amazon, стояло больше всего положительных (как мы помним, оценки 4,5) отзывов, что свидетельствует о любви читателей к определённому автору в рамках рассматриваемого набора данных.

3. Название третье главы

Требуется четко и подробно, шаг за шагом описать, что вы и как сделали. Содержание данного раздела должно быть таким, чтобы любой человек мог шаг за шагом повторить ваши результаты.

4. Анализ результатов

После опытов следует анализ полученных результатов (графиков, таблиц и др.), обсуждение полученного, выводы.

5. Заключение

Заключение должно содержать основные результаты работы, сделанные вами выводы и выдвинутые гипотезы.

Список литературы / References

[1] Сергей Орлов, “Журнал о современных тенологиях”, *КОМПЬЮТЕРРА*, 16 января 2019.

[2] Rosie Murphy, “Local consumer review survey 2018”, *Веб-сайт*, 7 декабря 2018.

[3] Medhat, Walaa and Hassan, Ahmed and Korashy, Hoda, “Sentiment analysis algorithms and applications: A survey”, *Ain Shams engineering journal*, **5** (2014), 1093–1113.

[4] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsuan-Shou, “Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news”, *Knowl-Based Syst*, **41** (2013), 89–97.

[5] Teng-Kai Fan, Chia-Hui Chang, “Blogger-Centric Contextual Advertising”, *Expert Systems with Applications*, **38**:3 (2011), 1777-1788.

[6] Gamal, Donia and Alfonse, Marco and M El-Horbaty, El-Sayed and M Salem, Abdel-Badeeh, “Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains”, *Machine Learning and Knowledge Extraction*, **1** (2019), 224–234.

[7] Sasikala P, L.Mary Immaculate Sheela, “Sentiment Analysis and Prediction of Online Reviews with Empty Ratings”, *International Journal of Applied Engineering Research*, **13** (2018), 11525–11531.

[8] Sasikala P, L.Mary Immaculate Sheela, “Sentiment Analysis of Online Food Reviews using Customer Ratings”, *International Journal of Pure and Applied Mathematics*, **119** (2018), 3509–3514.

[9] Bhatt, Aashutosh and Patel, Ankit and Chheda, Harsh and Gawande, Kiran, "Amazon Review Classification and Sentiment Analysis", *International Journal of Computer Science and Information Technologies*, **6** (2015), 5107–5110.

[10] McClure, Nick, "TensorFlow machine learning cookbook", 2017, 185–187.

[11] Müller, Andreas C and Guido, Sarah and others, "Introduction to machine learning with Python: a guide for data scientists", 2016, 339–341.

[12] .

[13] http://www.machinelearning.ru/wiki/index.php?title=.