Санкт-Петербургский государственный университет

МИРОНОВ Андрей Александрович

Выпускная квалификационная работа

Нейросетевые подходы выявления проблем маркетинговой стратегии компаний на основе анализа пользовательских дискуссий в социальных сетях

Уровень образования: магистратура

Направление 01.04.02 «Прикладная математика и информатика»

Основная образовательная программа ВМ.5505.2019 «Математическое и информационное обеспечение экономической деятельности»

Научный руководитель:
доцент кафедры Технологии программирования
кандидат технических наук
Блеканов Иван Станиславович

Рецензент:
Начальник отдела стратегических инициатив,
центр разработки и монетизации данных
ООО “Газпромнефть-Цифровые решения”
Кононов Ярослав Сергеевич

Санкт-Петербург

2021

Содержание

1 Введение

1.1 Актуальность работы

Социальные сети прочно вошли в повседневную жизнь и теперь очень многие события и явления нашей жизни так или иначе связаны с ними. С помощью анализа социальных сетей можно узнать очень многое об их пользователях, об их мнениях, предпочтениях и реакции на то или иное явление или событие. Такой анализ возможен по причине того, что в различных соцсетях сейчас присутствует более половины населения Земли.

Согласно последним данным население Земли составляет около 7,75 млрд. человек. Из них около 4.54 млрд являются пользователями интернета. Из этих 4.54 млрд. около 3.8 млрд являются активными пользователями различных социальных сетей, проводя в них в среднем около 2 часов 24 минут в день.[22]

1.2 Практическая значимость работы

Благодаря такому охвату, социальные сети становятся важным фактором для многих сфер жизни, в том числе и для бизнеса. Существует множество исследований на тему влияния социальных сетей на отельный[7], рекламный бизнес[14], на производство одежды[11] и др. Известные случаи когда после публикаций в “Twitter” акции компании Илона Маска падали на 10% [1] также можно отнести к влиянию социальных сетей на состояние фондового рынка. Очевидно, что объём информации имеющий такое влияние на все сферы жизни должен заинтересовать исследователей, специализирующихся на анализе мнений и настроений в обществе. В основном подобные исследования представляют собой изыскания на тему “анализа мнений”(*opinion mining*) или анализа тематик. Такую возможность исследователям предоставляют сами пользователи, активно делящиеся в соцсетях своим мнением по тому или иному вопросу.

Поэтому довольно интересной становится задача оценки привлекательности продукта для пользователя с использованием информации, полученной из социальных сетей.

1.3 Цель работы

Целью работы является создание механизма, позволяющего производить анализ и выявлять проблемы в маркетинговой стратегии компаний на основе пользовательских публикаций в социальных сетях(в частности в “Twitter”).

1.4 Задачи работы

Однако поиск решения для такой задачи сопряжён с рядом сложностей. Основная из которых - большая “зашумлённость” данных, т.к мы работаем с пользовательским контентом, который подразумевает ошибки, сленговые выражения, рекламные сообщения, выдаваемые за пользовательские и другие виды информационного “шума”. Также сложности добавляют краткость контента(пользователи реже развёрнуто выражают свою мысль, чаще используя короткую форму, а некоторые соцсети, например “Twitter” намеренно ограничивают длину одного сообщения для пользователя) и его мультиязычность(проблема особенно характерна для анализа глобального явления или продукта, выпускаемого транснациональной корпорацией, мнение о котором могут публиковать представители разных стран).

Ещё более сложной задачей является не просто оценка удовлетворённости пользователей некоторым продуктом, а оценка удовлетворённости пользователя конкретными аспектами данного продукта, так называемый Aspect-Based Sentiment Analysis. Он позволяет оценить отношение пользователя к некоторому конкретному аспекту продукта и может быть полезен в случае, если потребитель удовлетворён одним аспектом, однако недоволен другим. Например, рассмотрим такой отзыв: “Phone’s sharge stands quite well, touch screen sensitivity is great, but it’s web browsing speed is not satisfactory”. Данный отзыв вполне может быть отнесён к позитивным, однако о скорости веб-браузера пользователь высказался в негативном ключе. Для того, чтобы извлечь такую, более точную информацию об отношении пользователя к продукту нужно определять тональность пользователя по отношению к каждому упомянутому им аспекту.

Поиску решения такой задачи и преодоления указанных сложностей посвящена в том числе и данная работа.

Для достижения данной цели необходимо решить следующие задачи:

- Анализ литературы по данной тематике(статьи, другие публикации)
- Анализ технологических решений, применяемых при решении схожих задач
- Разработка архитектуры
- Разработка методов
- Тестирование
- Оценка качества решения

В конечном итоге результатом данного исследования является механизм, позволяющий производить мониторинг трендов, появляющихся в социальных сетях и интересующих ту или иную бизнес-структуру. Кроме того он позволит мониторить удовлетворённость пользователей продуктом, причём удовлетворённость конкретными аспектами продукта, которые интересуют производителя.

2 Обзор существующих методов и инструментов анализа социальных сетей

2.1 Обзор существующих методов

Aspect Based Sentiment Analysis(далее "ABSA") представляет собой довольно обширную область так называемого *NLP(Natural Language Processing)*, основу которой составляют методы, направленные на выявление тональности текста относительно каждого конкретного аспекта, упоминаемого в тексте.

Основные принципы и методы анализа настроений пользователей в социальных сетях довольно полно описаны в книге *"Sentiment Analysis in Social Networks"*[18]. Однако данное издание не покрывает частный случай анализа тональности мнения по конкретным аспектам(*"Aspect Based Sentiment Analysis"*) Ранние публикации в этой области в основном используют ручную разметку текста по аспектам и их тональности и затем классификацию на основе модели, обученной на таких размеченный вручную данных с помощью, например, SVM[19], [9]. Если изучить более поздние публикации на эту тему, то методы, используемые в исследованиях можно ословно разделить на 3 группы:

- Lexicon-based approaches - основанные исключительно на синтаксических и морфологических правилах.
- Machine learning approaches - основанные на методах машинного обучения
- Hybrid approaches - соединение синтаксических методов с методами на основе машинного обучения.

Lexicon-based approaches отличаются тем, что не требуют никакой обучающей выборки, т.к используют уже готовые базы тональностей слов, такие как WordNet [12] или SentiWordNet [4]. Недостаток таких методов в том, что они не учитывают соответствие слов-маркеров тональности словам-объектам.

Machine learning approaches - в свою очередь также могут быть разделены на 2 подгруппы:

- Supervise, т.е использующие "обучение с учителем"
- Unsupervised, т.е использующие "обучение без учителя"

Первая подгруппа является наиболее обширной. К ней можно отнести подходы с использованием SVM(Support Vector Machine), свёрточных нейронных сетей[20], трансферных

нейронных сетей[17]. В целом такие методы показывают довольно высокие результаты [15] [16], однако к их недостаткам можно отнести тот факт, что они сильно зависят от наличия большого обучающего датасета. Также подобные методы сильно зависят от качества извлечения аспектов. Unsipervised методы как правило используются только в связке с другими подходами, например как в [8].

Наибольшую долю в подобных исследованиях занимают так называемые *hybrid approaches*, т.е использующие и машинное обучение и синтаксические методы, например [8], [3].

Основным игроком на российском рынке услуг по оценке информационного фона того или иного лица или события в данный момент является компания “Медиалогия”. По подписке компания предоставляет сервис, позволяющий оценить “вес” бренда в соцсетях. Оценка даётся на основании количества и качества упоминаний бренда в публикациях в различных социальных сетях. Так называемый “СМ индекс” зависит от аудитории автора или сообщества, которые опубликовали сообщение, аудитории их репостеров, и от значения вовлечённости сообщения (количество лайков, репостов, комментариев). Оценка учитывает тип сообщения (пост/репост/комментарий), а также тип площадки (персональный аккаунт, группа или сообщество). На мировом рынке лидирующую позицию занимает “Youscope”, предоставляющий аналогичные услуги, но уже в мировом масштабе.

2.2 Обзор используемых инструментов

Для обработки данных и построения нейронной сети мной был выбран язык программирования Python версии 3.7 и интерактивная среда разработки Google Colab (RAM: 12 GB), т.к благодаря большому количеству компонентов для машинного обучения и построения нейросетей данная комбинация является наиболее популярной в области академических исследований нейронных сетей.

Для создания и обучения нейронных сетей как правило используются фреймворки, которые позволяют разработчику не обращать внимания на многие низкоуровневые аспекты конструирования моделей для обучения и вместо этого сосредоточиться на конкретных параметрах, определяющих архитектуру и метод обучения нейросети. Для данной работы была выбрана библиотека TensorFlow от компании Google, и её реализация для языка программирования Python.

В качестве инструмента для синтаксического анализа мной был выбран Stanford Parser[5], т.к в данный момент он является наиболее актуальным и показывающим наилучшие результаты компонентом для анализа текста на предмет логической связи между словами, выде-

ления частей речи итд. Используется последняя на момент создания данной работы версия парсера - 4.2.0

В отличии от большинства работ, посвящённых данной теме, использующих наиболее популярный способ представления текстовой информации в виде векторов, а именно word2vec [21], мной для данной работы был выбран *Universal Sentence Encoder(USE)*.

Причина такого выбора в том, что USE позволяет создавать векторное представление для текста независимо от того, на каком языке этот текст написан. Благодаря этому векторное представление слова “молоко”будет полностью идентично представлению слова “milk что является важным критерием для данной работы, т.к здесь производится анализ пользовательских публикаций в соцсетях, которые могут быть сделаны на различных языках.

Ещё одним преимуществом USE перед word2vec является возможность создать векторное представления сразу для целого предложения или выражения, которое также будет отражать его смысл. Т.е вектора для слова “weekend”и для фразы “Saturday and Sunday”будут находиться на минимальном расстоянии друг от друга.

2.3 Обзор метрик оценки качества

Точность работы нейронной сети оценивалась с помощью нескольких метрик, таких как *Accuracy, Precision, Recall и F-Measure*. Объясним их значение более подробно. Accuracy(простая точность) - наиболее понятная и очевидная метрика из данного списка. Представляет собой простое отношение верно угаданных результатов к размеру всего датасета, т.е определяется по формуле:

$$Accuracy = \frac{P}{N},$$

где P - количество верных результатов(в нашем случае - количество твитов с вернно угаданными, а N - общий размер выборки.

Precision(точность) представляет собой более сложную метрику и определяется как доля документов действительно принадлежащих данному классу относительно всех документов которые система отнесла к этому классу. Это значение определяется формулой:

$$Precision = \frac{TP}{TP + FP},$$

где TP - количество твитов, верно отнесённых моделью к соответствующему классу(истинно-положительных решений), а FP - количество твитов, неверно отнесённых моделью к тому же классу(ложно-положительных). Далее рассмотрим метрику Recall(полнота). Она определя-

ется формулой:

$$\textit{Recall} = \frac{TP}{TP + FN},$$

где FN - ложно-отрицательные решения, т.е количество случаев, в которых твит, принадлежащий некоторому классу, не был помечено таковым нейросетью. Наконец, F-Measure представляет собой гармоническое среднее значений Precision и Recall, что позволяет найти наилучший алгоритм с учётом обоих значений данных метрик. В нашем случае Precision и Recall являются равнозначными(т.е одна не имеет приоритета над другой) и вычисляются как

$$F = 2\frac{Precision \times Recall}{Precision + Recall}$$

Т.к в данной работе рассматривается классификация твитов(т.е задача отнесения твитов к тому или иному классу), при том, что в нашей задаче необходимо распределить все твиты между 12 классами, то Precision и Recall будем определять как

$$\textit{Precision}_{\textit{micro}-\textit{average}} = \frac{\sum\limits_{n} TP_n}{\sum\limits_{n} TP_n + \sum\limits_{n} FP_n}$$

$$\textit{Recall}_{\textit{micro}-\textit{average}} = \frac{\sum\limits_{n} TP_n}{\sum\limits_{n} TP_n + \sum\limits_{n} FN_n},$$

где TP_n означает количество твитов, верно отнесённых моделью к классу n, FP_n - количество твитов, неверно отнесённых к классу n, а FN_n - количество твитов, принадлежащих классу n, которые не были помечены таковыми моделью.

Использование данного набора метрик является стандартным в работах, посвящённых машинному обучению, схожие расчёты могут быть замечены во многих работах, посвящённых данной теме[9][8][3].

Для вычисления значений данных метрик перед началом обучения были выделены 1000 случайных твитов, которые были удалены из обучающей выборки, чтобы при тестировании система встретила их впервые, и обозначены как *тестовый датасет*. Далее на всех этапах обучения показатели качества работы нейросети определялись на основе результатов, показанных на этой тестовой выборке.

3 Разработка инструментов и методов для выявления проблем маркетинговой стратегии компаний на основе анализа пользовательских дискуссий в социальных сетях

3.1 Разработка нейросетевого подхода

Для решения данной задачи было решено использовать рекуррентные нейронные сети. Примерная схема изображена на рис.1.

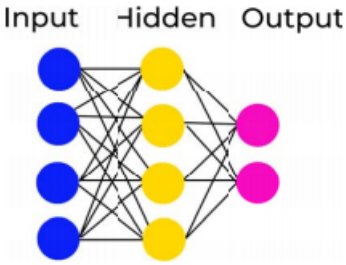

Рис. 1 Примерная схема нейронной сети.

На вход нейронной сети подаются твиты, преобразованные в вектора с помощью Universal Sentence Encoder-a. Соотношение между значениями входных и выходных нейронов можно проиллюстрировать следующей формулой:

$$Y_i = g(\sum_j W_{ij} * a_j), \tag{1}$$

где a_j ссылается на входной параметр, W_{ij} - матрица весов, а g - некоторая функция активации. В данном случае функцией активации был выбран сигмоид:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

Обучение такой сети происходит с помощью последовательной модификации матрицы весов. Такая модификация осуществляется путём сравнения получившихся значений с целевыми и корректировки соответствующего элемента матрицы весов с помощью метода обратного распространения ошибки(back propagation methods). Сеть состоит из 4-х полностью соединённых слоёв(Dense layers). Все нейроны такого слоя соединены со всеми нейронами следующего за ним слоя.

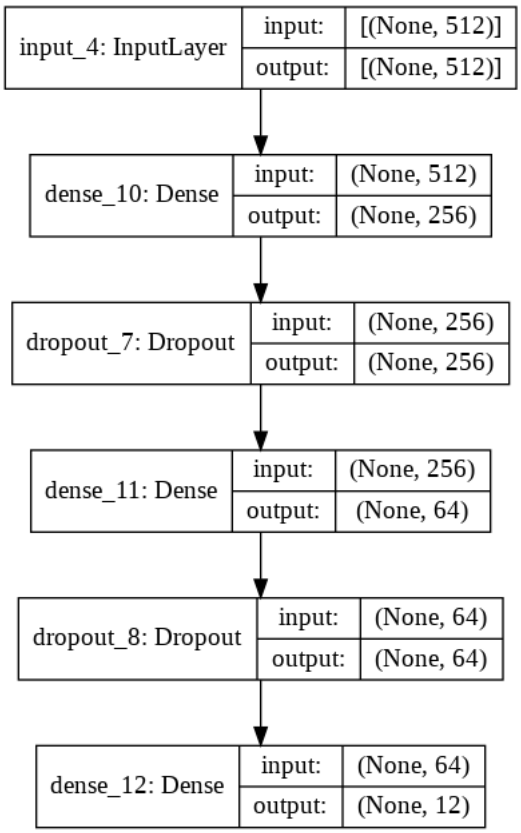

Рис. 2 Архитектура нейронной сети для классификации аспектов.

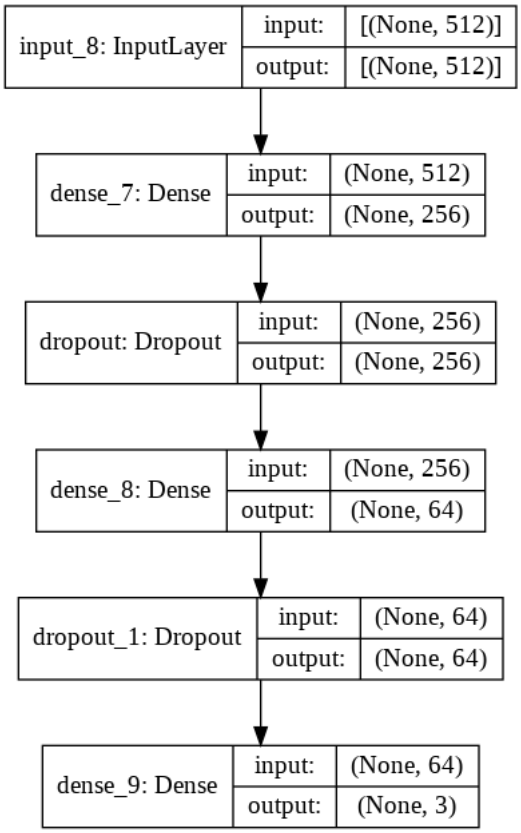

Рис. 3 Архитектура нейронной сети для анализа тональности.

На входной слой подаётся вектор размерности 512, получаемый с помощью Universal Sentence Encoder-a. Первый слой является полностью соединённым(Dense layer) с применяемой к нему сигмоидной функцией активации(2).

Следующие две группы представляют собой последовательности Dense и Dropout слоёв. Dropout слой применяется с целью предотвратить так называемое “переобучение”. Это достигается тем, что на этапе Dropout-a некоторое количество нейронов “отключаются”. В данном случае отключаются 20% нейронов.

В результате работы нейросети каждому твиту сопоставляются один или несколько аспектов, которые в нём упоминаются. В данном случае аспект считался упомянутым в твите, если значение, соответствующее аспекту на выходе нейросети превышало 0,7.

3.2 Препроцессинг данных, выделение аспектов

Algorithm 1 Предварительная обработка датасета

$tweetsForAnalysis \leftarrow []$
for all $tweets$ **do**
    if $language(tweet) = english$ **then**
        if $numOfBrandsInTweet(tweet) = 1$ & $numOfHashtags(tweet) \geq 1$ & $numOfMentionedBrands(tweet) = 1$ **then**
            $tweetsForAnalysis.append(tweet)$
        end if
    end if
end for
$tweetsByAuthor \leftarrow arrangeTweetsByAuthor(tweetsForAnalysis)$
for all $tweetsByAuthor$ **do**
    if author has 2 or more equal tweets **then**
        remove all tweets of this author
    else
        $resultTweets.append(alltweetsofauthor)$
    end if
end for
return $resultTweets$

На первом этапе датасет очищается от всех сообщений, кроме тех, которые содержат мнение об одном и только об одном из брендов. В случае твитов, например, оставляются

только содержащие ровно один хэштег, соответствующий рассматриваемому бренду.

Затем датасет очищается от лишнего “шума” и спама. Здесь под спамом подразумеваем большое количество одинаковых публикаций, сделанных одним и тем же пользователем и, очевидно, не подходящих нам для исследования проводимого в данной работе. Для этого было сделано предположение, что если пользователь не является рекламным “ботом”, то мы, скорее всего, не увидим у него двух одинаковых сообщений. Поэтому были удалены сообщения от всех пользователей, которые имели хотя бы 2 одинаковые публикации. Псевдокод представлен на Алгоритме 1.

Следующей задачей стало выделение основных аспектов(*aspect mining*), отношение к которым и является информацией, с помощью которой будет произведён анализ маркетинговой стратегии брендов. Для выделения аспектов мной был применён(с небольшими изменениями) метод, описанный в [8], основанный на разбиении всех твитов на так называемые *NP-chunks*. NP-chunks(noun phrase chunks) представляют собой короткие фразы, на которые разбивается каждое предложение на основе лингвистических правил, причём каждая такая фраза должна содержать имя существительное. Т.е например из предложения “the little yellow dog barked at the cat”будут извлечены фразы “the little yellow dog”и “the cat”. Такое разбиение позволяет получить информацию обо всех сущностях, упоминаемых в предложении. Для задачи выделения NP-chunks из каждого твита был использован Stanford NLP Parser.

Из получившегося множества фраз с помощью метода, описанного в [8] были выделены слова и фразы, которые потенциально могут быть отнесены к аспектам. А именно: если перед именем существительным было найдено прилагательное, не относящееся к разряду качественных(для определения таковых была использована база opinion lexicon из пакета NLTK), то считаем эту связку прилагательное-существительное потенциальным аспектом, иначе - считаем потенциальным аспектом только имя существительное.

Все слова и фразы, собранные таким образом были собраны и преобразованы в вектора с помощью “Universal Sentence Encoder”. Затем получившиеся вектора были разбиты на кластеры с помощью метода К-средних. Оптимальное количество кластеров вычислялось “методом силуэта который можно представить как:

$$s_i = \frac{b(i) - a(i)}{max(b(i), a(i)}, \tag{3}$$

где a(i) - среднее расстояние между объектами i-го кластера, b(i) - среднее расстояние от i-го до ближайшего к i-му кластера. Такой коэффициент вычисляется для каждого кластера в данном разбиении и чем выше среднее значение коэффициента s_i для данного разбиения, тем более точным является само разбиение.

После разбиения на кластеры для каждого кластера было найдено слово или словосочетание, векторное представление для которого наиболее близко к центроиду кластера. После чего в результате эмпирического анализа получившихся кластеров выделяются основные аспекты, о которых пользователи высказывались в своих публикациях.

3.3 Обучение модели для выделения аспектов

Для обучения модели, выделяющей аспекты использовалась обыкновенная рекурентная нейронная сеть. Примерная схема такой нейросети изображена на рис.2. Обучающая выборка для указанной сети была построена следующим образом: из каждого твита были выделены NP-Chunks, описанные выше, затем эти np-chunks были преобразованы в векторную форму с помощью “Universal Sentence Encoder”. После чего была вычислена мера подобия между указанными векторами и векторными представлениями центроид кластеров, описанных в предыдущем пункте. Значение подобия вычислялось как мера косинусного сходства:

$$cos\_similarity(a,b) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{4}$$

Таким образом была сформирована обучающая выборка для обучения модели, извлекающей аспекты из твитов.

Обучение производилось на протяжении 10 эпох, данные на вход нейросети подавались батчами по 10 np-chunks в каждом батче. В качестве функции оптимизации использовался оптимизатор Adam[6] со значением learning rate 0,01.

3.4 Обучение модели для анализа тональности

Схема модели, применявшейся для анализа тональности публикации по отношению к аспекту, изображена на рис.3. При создании обучающей выборки твиты были разбиты на уже упомянутые выше NP-chunks. затем для каждого аспекта(они уже соотнесены с твитами на предыдущем шаге) были найдены NP-chunks, в которых упоминается этот и только этот аспект(т.е не упоминаются другие аспекты из нашего списка). Далее все эти NP-chunks были проанализированы на предмет их тональности с помощью Stanford NLP Parser, после чего методом ”голосования”данному аспекту назначалось значение тональности. Каждому аспекту сопоставлялся вектор его тональности размерности 3, в котором первая компонента соответствовала положительной тональности, вторая - отрицательной, а третья - нейтральной.

Другими словами если аспект был помечен, как упомянутый положительно, то его вектор имел вид [1, 0, 0].

Обучение для этой сети также производилось на протяжении 10 эпох, в качестве функции оптимизации также была использована функция Adam с learning rate 0,01.

4 Тестирование и оценка качества разработанного решения

4.1 Постановка эксперимента

В эксперименте оценивались качество работы следующих частей механизма:

- Выявление аспектов
- Анализ тональности

Для получения результатов эксперимента необходимо:

- Получить оценку качества для выявления аспектов
- Получить оценку качества для выявления тональности по отношению к аспектам
- Получить общую оценку качества
- Получить оценку мультиязычности механизма

4.2 Описание датасета

Датасет состоит из публикаций, сделанных в социальной сети “Twitter”с 01.12.2016 по 01.09.2017 на различных языках. Сбор производился с помощью поискового робота (web crawler), который основан на REST-запросах. Всего в датасете содержатся 6 519 273 твитов. Датасет предоставлен мне научным руководителем Блекановым Иваном Станиславовичем.

С помощью библиотеки spaCY для всех публикаций был определён язык. Распределение языков в датасете показано на рис.4.

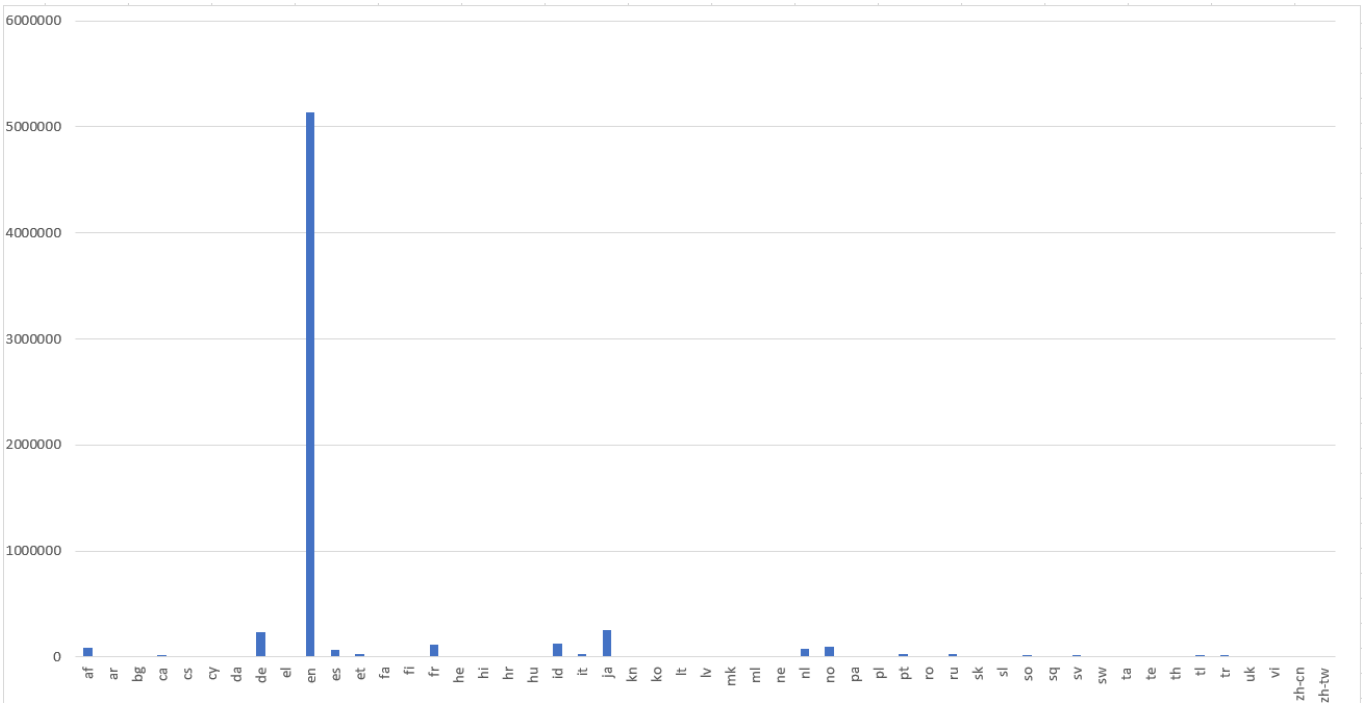

Рис. 4 Распределение твитов в коллекции по языкам.

Как видно на графике, подавляющее большинство твитов в коллекции относятся к англоязычным.

Также сразу можно выделить твиты, которые явно содержат мнение только о продукте одного бренда. Для этого необходимо выделить публикации, хэштеги которых содержат название только одного из брендов. В качестве брендов были взяты 20 основных присутствующих на этом рынке брендов[10] , а именно:

- ’adidas’
- ’asics’
- ’balenciaga’
- ’calvinklein’
- ’converse’
- ’dc’
- ’fila’
- ’jordan’
- ’lacoste’

- ’merell’
- ’new balance’
- ’nike’
- ’puma’
- ’reebok’
- ’saucony’
- ’skechers’
- ’timberland’
- ’tommy hilfiger’
- ’underarmour’
- ’vans’

Распределение таких твитов по брендам на рис.5

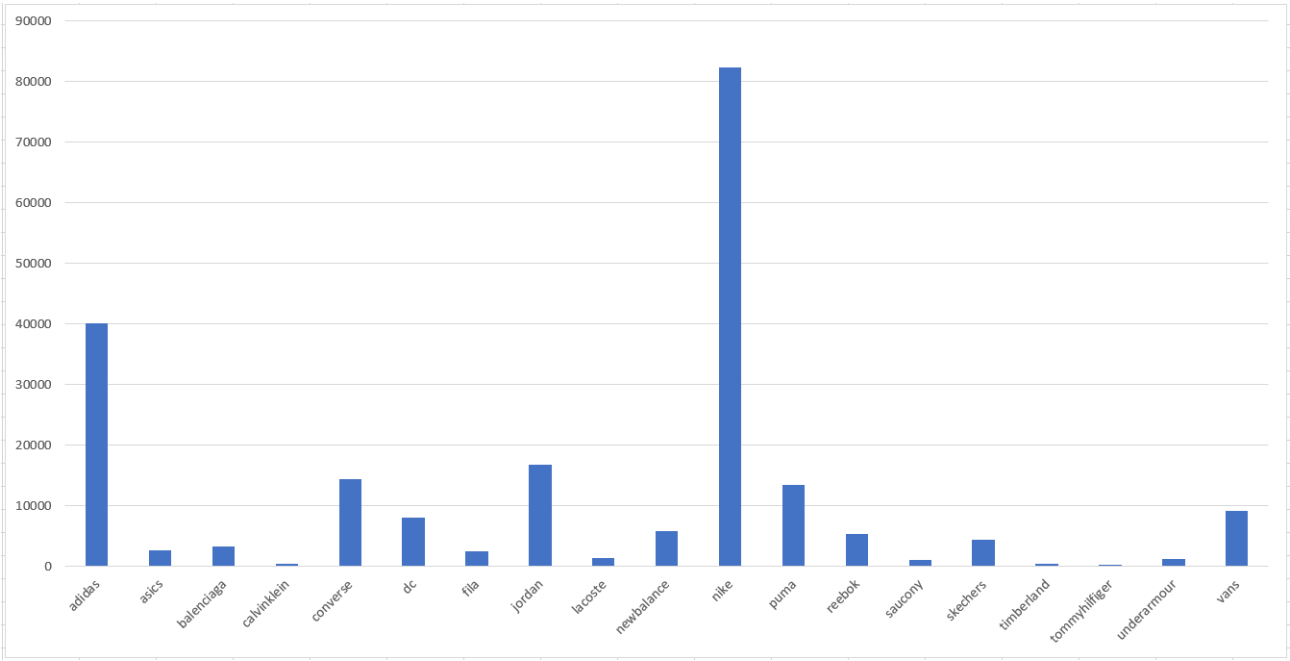

Рис. 5 Распределение твитов в коллекции по брендам.

4.3 Результаты обучения и анализа для задачи выделения аспектов

Для проверки механизма обучения были взяты данные из датасета, описанного в п. 3.2, над ними проведены операции по подготовке данных, описанные в п. 2.2.3. Процесс обучения на протяжении 10 эпох представлен на рис. 6.

Таблица 1: Результаты обучения для классификатора аспектов

Accuracy	Precision	Recall	F1-measure
88%	86%	84%	85%

Все значения в смысле, обозначенном в п. 2.1.1

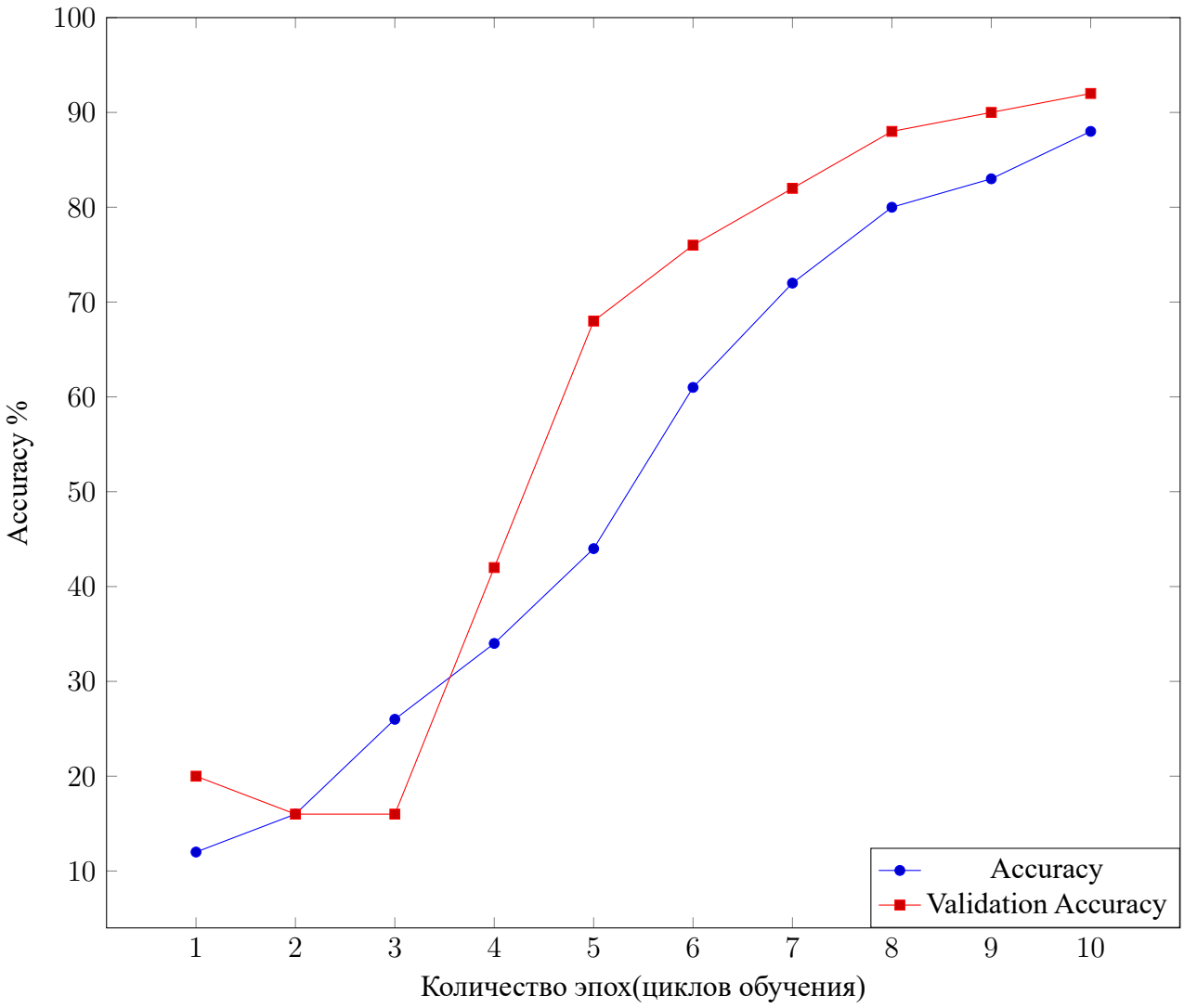

Рис.6. Обучение на протяжении 10 эпох

В Таблице 1 можно увидеть довольно высокие результаты, что говорит об успешном

обучении части общего механизма, отвечающей за выделение аспектов из твитов.

4.4 Результаты обучения и анализа для задачи анализа тональности

После 10 эпох обучения для модели, анализирующей тональность аспектов, были получены следующие результаты:

Таблица 2: Результаты обучения для классификатора тональности аспектов

Accuracy	Precision	Recall	F1-measure
91%	88%	83%	85%

Все значения в смысле, обозначенном в п. 2.1.1

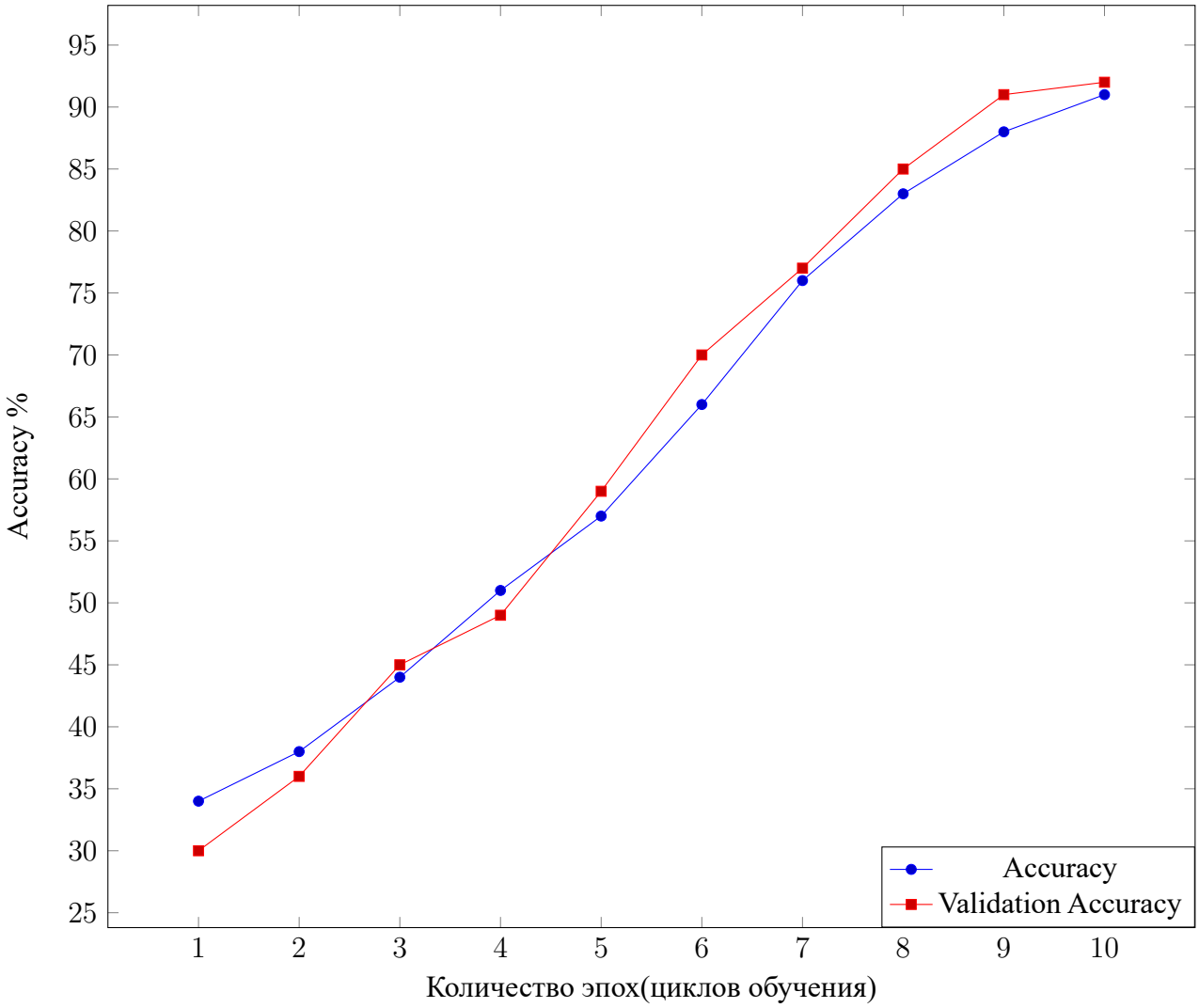

Рис.7. Обучение на протяжении 10 эпох

Модель, классифицирующая аспекты твитов по их тональности показывает также довольно высокие результаты, что говорит об успешности применяемого подхода к решению задачи Aspect-Based Sentiment Analysis.

4.5 Проверка мультиязычности

Как было заявлено в п.2, разработанный механизм помимо прочего должен также обладать свойством мультиязычности, т.е должен одинаково качественно работать для твитов на разных языках. Для проверки из выборки были взяты твиты на русском языке и отфильтрованы согласно принципам, изложенным в п. 2.2.3. Из отобранных твитов случайным образом было выделено 200 твитов, которые были размечены вручную. Выбор именно русского языка обусловлен двумя причинами. Первая: для автора работы русский является родным и он может с наилучшей точностью создать размеченную выборку именно из твитов на этом языке. Вторая: русский и английский являются языками из разных языковых групп, что позволит проверить качество работы алгоритма на языке с другими принципами построения предложений.

Твиты были размечены и проанализированы моделью, обученной в предыдущем пункте. Для деления твитов на NP-Chunks использовалось расширение для Stanford NLP Parser, позволяющее работать с русскоязычными твитами. Результаты:

Таблица 3: Результаты для классификатора аспектов в случае анализа русскоязычных твитов

Accuracy	Precision	Recall	F1-measure
85%	83%	81%	82%

Все значения в смысле, обозначенном в п. 2.1.1

Таблица 4: Результаты для классификатора тональности в случае анализа русскоязычных твитов

Accuracy	Precision	Recall	F1-measure
88%	87%	84%	85%

Все значения в смысле, обозначенном в п. 2.1.1

Таким образом мы можем заключить, что и для русского языка данный механизм показывает довольно высокие результаты, что позволяет нам говорить о его мультиязычности.

4.6 Общие результаты тестирования на реальном кейсе

Чтобы продемонстрировать возможности механизма можно рассмотреть результаты его работы для некоторого массива реальных данных. Для этого были проанализированы твиты, опубликованные в период с декабря 2016 по июль 2017 года и на основе данных, полученных из нейросети, была составлена так называемая *тепловая карта*.

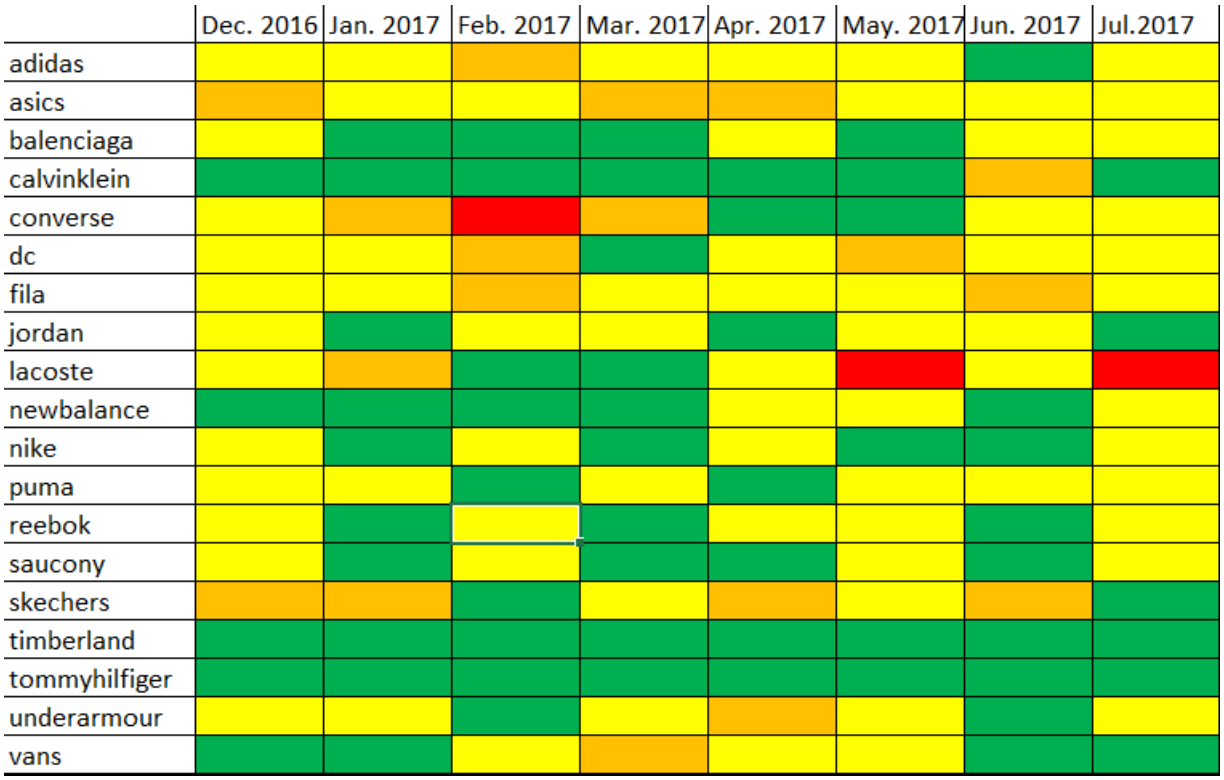

Рис. 8 Результат работы алгоритма для реальных данных в виде тепловой карты.

Здесь ячейка зелёного цвета обозначает, что в данный промежуток времени доля положительно оценённых аспектов в публикациях пользователей составила от 75% до 100%, жёлтый цвет – 50% - 75%, оранжевый цвет – 25% - 50% и красный цвет ячейки – до 25% положительно оценённых аспектов.

Теперь рассмотрим таблицу более подробно. Допустим, пользователя заинтересовали причины, по которым у бренда “lacoste” в мае 2017 года большое количество отрицательных отзывов. Тогда строим график по долям отзывов в данный период по всем аспектам.

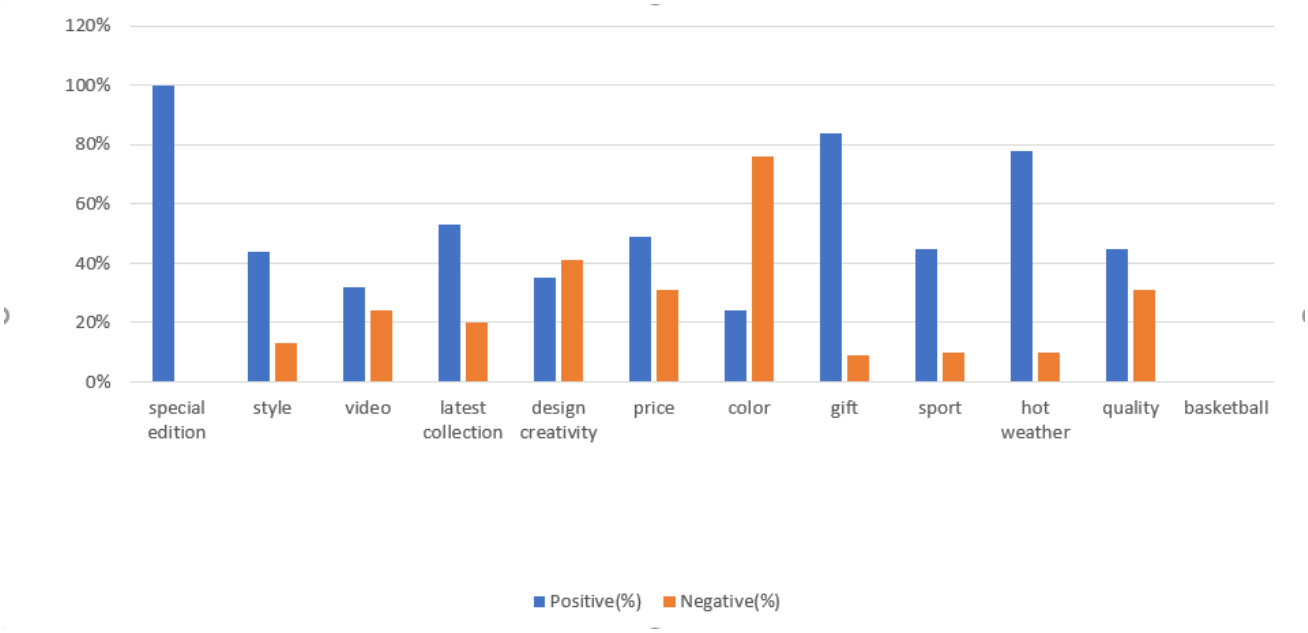

Рис. 9 Результат работы алгоритма для реальных данных в виде графика удовлетворённости аспектами бренда lacoste в мае 2017 года.

Здесь видно, что, судя по всему, в этом месяце пользователей не устроили цвет(доля отрицательных отзывов на аспект color очень высока) и дизайн в целом. С помощью алгоритма такую информацию можно получить для каждого бренда и для любого промежутка времени.

5 Заключение

5.1 Результаты работы

В ходе работы был разработан механизм, позволяющий анализировать большие объёмы пользовательских публикаций на тему какого-либо товара с целью анализа маркетинговой стратегии компаний в частности и мониторинга отношения пользователей к брендам, находящимся в данный момент на рынке в целом. Для решения данной задачи были применены нейросетевые методы анализа текстовой информации.

Механизм был обучен и протестирован на реальных данных(публикациях, сделанных в социальной сети “Twitter”), относящихся к теме кросовок. Результаты тестирования можно назвать успешными, т.к обе модели, присутствующие в алгоритме, показали довольно высокую точность. Также, благодаря использованию Universe Sentence Encoder-а была решена задача анализа публикаций на различных языках, что в ходе работы было также успешно протестировано.

Несмотря на привязку данного исследования к датасету, содержащему публикации по теме “sneakers”, в алгоритм в п.2 описан в общем виде, без каких-либо привязок к конкретным массивам данных, что означает, что он может быть применён и для анализа ситуации на любом другом рынке, где имеются пользователи, активно публикующие своё мнение о продуктах в социальных сетях.

5.2 Перспективы развития

Участком, требующим более детальной проработки остаётся определение аспектов, для которых впоследствии проводится анализ, т.к в данной работе аспекты были выделены из общего списка потенциальных аспектов эмпирическим путём. Это можно назвать возможной точкой приложения усилий для дальнейшего совершенствования алгоритма.

Однако в целом все задачи, поставленные в п.1 были выполнены в полном объёме.

6 Список литературы

Список литературы

[1] Медиалогия: Технологии[Электронный ресурс] – Режим доступа: https://www.mlg.ru/about/technologies/ – Загл. с экрана (09.05.2021)

[2] РБК: Илон Маск назвал акции Tesla слишком дорогими[Электронный ресурс] – Режим доступа:https://www.rbc.ru/business/01/05/2020/5eac7ea19a79478013cebe8a, свободный – Загл. с экрана(10.05.2021)

[3] Aniket Mukherjee. Aspect Based Sentiment Analysis of Student Housing Reviews / Aniket Mukherjee, Shiv Jethi, Akshat Jain, Ankit Mundra // In Proceedings of the 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)

[4] Baccianella S. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. / Baccianella S, Esuli A, Sebastiani F – Режим доступа: http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf, свободный – Загл. с экрана(10.05.2021)

[5] Cristopher D. Manning. The Stanford CoreNLP Natural Language Processing Toolkit / Cristopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel // In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System DemonstrationsAt: Baltimore, Maryland.

[6] Diederik P. Kingma. Adam: A Method for Stochastic Optimization / Diederik P. Kingma, Jimmy Ba // In Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015

[7] Firas Mohamad Halawani The Effect of Social Media on Hotels' Business Performance in the Lebanese Hotel Sector: Effect of Social Media on Hotels' Business Performance / Patrick C.H. Soh, Saravanan Muthaiyah // Journal of Electronic Commerce in Organizations (JECO), 2019, 17, 3, 54.

[8] Ganpat Singh Chauhan A two-step hybrid unsupervised model with attention mechanism for aspect extraction / Ganpat Singh Chauhan, Yogesh Kumar Meena, Dinesh Gopalani, Ravi Nahta // Expert Systems with Applications Volume 161, 15 December 2020

[9] J. Wagner DCU: Aspect-based polarity classification for SemEval task 4 / P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, and L. Tounsi // In Proc. 8th Int. Workshop Semantic Eval. (SEMEVAL), 2014, pp. 223–229.

[10] John Motson. Global $88 Bn Sneakers Market to 2024 // GlobeNewswire, COMTEX News Network, Inc., 2019, pp. 45-54

[11] Lee, Jung Eun Visual communication of luxury fashion brands on social media: effects of visual complexity and brand familiarity / Hur, Songyee, Watkins, Brandi // Journal of brand management; Sep 2018, 25 5, p449-p462, 14p.

[12] Miller GA Wordnet: a lexical database for english[Электронный ресурс] – Режим доступа: https://doi.org/10.1145/219717.219748(Дата обращения 10.05.2021)

[13] Nadezhda Chechneva, Simple and Efficient Approach to the Aspect Extraction from Customers'Product Reviews // In Proceedings of the 26th conference of Fruct association, 2020, pp. 210-217

[14] Noguti Motivations to use social media: effects on the perceived informativeness, entertainment, and intrusiveness of paid mobile advertising / Valeria, Waller, David S., // Journal of marketing management; Oct. 12 2020, 36 15-16, p1527-p1555, 29p.

[15] Pang B Thumbs up?: sentiment classification using machine learning techniques / Lee L, Vaithyanathan S // In: Proceedings of the ACL-02 conference on empirical methods in natural language processing - volume 10, EMNLP '02. Association for Computational Linguistics, Stroudsburg, pp 79–86

[16] Prabowo R Sentiment analysis: a combined approach[Электронный ресурс] / Prabowo R, Thelwall M – Режим доступа: https://doi.org/10.1016/j.joi.2009.01.003, свободный, Загл. с экрана(08.05.2021)

[17] Priyambada Ambastha Incident Detection From Social Media Targeting Indian Traffic Scenario Using Transfer Learning / Priyambada Ambastha, Maunendra Sankar Desarkar // In Proceedings of 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)

[18] Sentiment Analysis in Social Networks / Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu – Morgan Kaufmann, 2017 – 284 c.

[19] S. Kiritchenko NRC-Canada2014: Detecting aspects and sentiment in customer reviews / X. Zhu, C. Cherry, and S. Mohammad // In Proc. Int. Workshop Semantic Eval., 2014, pp. 437–442.

[20] Thet Naing Tun Intent Classification on Myanmar Social Media Data in Telecommunication Domain Using Convolutional Neural Network and Word2Vec / Thet Naing Tun, Khin Mar Soe // In Proceedings of the 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques , 2020, pp 312-321

[21] Tomas Mikolov. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I.Sutskever, K.Chen, G.Corrado, J.Dean // In Proceedings of NIPS, 2013

[22] We are social Ltd. official site[Электронный ресурс] – Режим доступа: https://wearesocial.com/digital-2020 – Загл. с экрана 21.04.2021