ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

###### **Факультет Гуманитарных наук**

**АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ТОНАЛЬНЫХ ФРЕЙМОВ**

Выпускная квалификационная работа

по направлению подготовки Фундаментальная и компьютерная лингвистика

образовательная программа «Компьютерная Лингвистика»

Руководитель:

доцент С.Ю.Толдова

Москва 2020

[1.Введение](#_9s02nph7fof4) 3

[*1.1. Извлечение мнений: основные задачи и подходы*](#_l3laibb3e8je) *4*

[*1.1.1 Основные задачи анализа тональности и способы их решения*](#_ozc9i18hb2te) *4*

[*1.1.2. Словари оценочной лексики. Методы автоматического составления тональных словарей*](#_2irz5dworqig) *5*

[*1.1.3. Подходы с использованием оценочных фреймов*](#_7zl80q4uis6t)6

[2.Ход работы](#_dn1mmr18nnvw) 7

[*2.2.Базовая проверка метода*](#_dlebfp77w1c1) *7*

[*2.1.1 Мультиклассовая классификация*](#_1wgr6n8tta30) *9*

[*2.1.2 Политематическая классификация*](#_fygxkvqj1pq6) *12*

[*2.1.3. Исследование предсказаний на неразмеченных данных*](#_b7yyk2x645dv) *13*

[*2.2.Основной метод*](#_gkspjfxgice) *14*

[*2.2.1 Политематическая классификация на примере DeepPavlov*](#_8zd6gyorp433)15

[3. Результаты](#_wl222kexnipw) 16

[4. Заключение](#_iq28of7r76f8) 16

[5. Список литературы](#_jcaocrfxrvd4) 16

[6. Приложение](#_mlhbntjpmltg) 18

# **1.Введение**

Тональный анализ очень востребованная и развивающаяся отрасль автоматической обработки естественного языка. Тональный анализ подразумевает под собой задачу автоматического извлечения настроения из текста. Каждому тексту, предложению или слову в зависимости от тональности которую он несёт присваивается определённое значение, полярность. Способы оценки полярности могут быть разными: бинарная [3], где отрицательная полярность это 0, а положительная 1, или шкала [4], где нижняя граница означает сильно отрицательную полярность, а верхняя - сильно положительную.

В наше время проводится множество исследований в этой области, в том числе и большими компаниями, так как тональный анализ становится жизненно важным в банковских системах, новостных агрегаторах, в клиентоориентированных сферах, работе политиков. Однако из-за специфики текстов, большая часть исследований узконаправленная [7][8]. Каждой сфере свойственна своя специфическая лексика, которая основана, на определённой теме: отзывы на фильмы или рестораны, тематические новости. Подобное разделение накладывает ограничение на дальнейшее использование получившихся моделей на других данных.

В этом исследовании мы основываемся на работе Лукашевич [1] и методе тональных фреймов, глаголо-опосредованной модели специфических связей между предикатом и его аргументами. В основе этого метода лежит идея, что предикат может влиять на полярность его аргументов (например “А выиграл Б”, следовательно у “А” есть положительный оттенок, а у “Б” - негативный).

В данной работе мы собираемся автоматически расширить существующий фреймовый тональный словарь, созданный и размеченный экспертами на основе данных собранных из Твиттера [5]. Таким образом целью является пополнение списка предикатов, которые могут влиять на общую оценку текста, а также в проверке гипотезы, что твиты, являясь атематичным, субъективным и эмоциональным материалом, могут дать более полную и точную базу для проведения эксперимента.

## *1.1. Извлечение мнений: основные задачи и подходы*

## *1.1.1 Основные задачи анализа тональности и способы их решения*

Тональный анализ - это задача извлечения мнения и субъективной информации из текста. Со времени появления первых работ в области [6], эта задача проделала долгий путь к высокотехнологичным нейросетевым моделям и алгоритмам, таким как BERT [9] (Bidirectional Encoder Representations from Transformers) или CSS (Contextual Semantic Search). Изначально это был алгоритм определения тональности текста, основанный на подсчёте определённых паттернов в тексте в совокупности с определением эмоционального состояния человека. С развитием компьютерных технологий, учёные начали выделять важные аспекты и параметры этой задачи: тематика, субъективность речи, объекты и субъекты речи.

Объективность текста - одна из основных проблем тонального анализа. Впервые её подняли в своей работе [10] Бо Панг и Лилиан Ли. Они сделали предположение, что эмоциональность и субъективность текста неразрывно связаны с полярностью. С помощью классификатора они разделили данные на субъективные и объективные и проанализировали получившийся результат, оказалось, что по субъективной информации получалась более точная оценка полярности, чем по объективной.

Основной метод решения задачи тонального анализа - классификация, алгоритм категоризации данных по классам. Для тонального анализа используется несколько основных вариантов классификации: бинарный (разбиение данных на две группы: негативную и позитивную), мультиклассовый (разбиение на несколько групп, где нижнее значение - крайне негативное, а верхнее значение - крайне положительное), линейный (где каждому объекту присваивается значение на определённом промежутке, где нижняя граница - это крайне негативное значение, а верхняя - крайне положительное). Обычно задачу классификации подразделяют на два способа решения[11]. Первая из них - тональный анализ документа, где каждому документу присваивается значение тональности. Вторая - тональный анализ клаузы, где текст разбивается на части в виде клауз или предложений, и каждой из частей присваивается некоторое значение тональности.

## *1.1.2. Словари оценочной лексики. Методы автоматического составления тональных словарей*

Другая важная часть тонального анализа - тональные словари [2], словари которые отражают полярность каждого слова. В основном словари разбиваются по тональностям (негативное, нейтральное или позитивное слово), однако, некоторые из них используют идею разбиения по размерностям. В EmoLex [14] слова разделены на три группы: Valence, Arousal, Dominance. Это отличается от классического бинарного разделения.полярностей или эмоционального разделения (злость, страх, грусть, радость). Этот подход может по новому осветить эту проблему и расширить тональные словари.

Есть несколько методов сбора тональных словарей. Первый - ручная разметка: с помощью краудсорсинга, исследователи собирают полярности слов. Таким образом данные разделяются между всеми респондентами и собирается полная и точная карта полярностей слов. Другой метод - Label Propagation [13], где размечается лишь небольшая часть данных, а остальные слова распределяются по классам алгоритмически.

Одной из важных проблем в тональных анализе является многозначность слов, в результате которой у слов в различных ситуациях может быть разная полярность. Например прилагательное “холодный” относительно звонков может быть в негативном ключе, а “холодный” относительно напитка в жару будет иметь позитивное значение. Для решения такой проблемы во многих работах используются тематические датасеты, где семантическое поле слов статично. Но такой подход не позволяет использовать полученную модель и результаты на других задачах и темах. Использование Твиттера как источника данных поможет нам избежать проблемы узконаправленности полученных данных, однако не решит проблему многозначности слов. Для данной задачи существует множество вариантов решения, одним из них является WordNet, который может разрешать подобные неоднозначности, используя большой объем данных.

## *1.1.3. Подходы с использованием оценочных фреймов*

Основная идея этой работы - улучшение существующих методов тонального анализа с помощью имплементации тональных фреймов, основанных на связи между предикатом субъектом и объектом глагола. В своей работе [12] Рашкин поднимает вопрос важности коннотативных фреймов. В ней он исследовал связи между субъектом и объектом и как предикат влияет на их полярность. Такой подход позволяет понять отношение автора к объектам речи. Одной из важных деталей такого подхода является тот факт, что предикат может влиять по разному на объект и субъект глагола в случае дачи оценки автором событию. Например, во фразе “А победил Б” “А” имеет положительную окраску в глазах автора высказывания, а “Б” наоборот негативную. Таким образом позитивная и негативная полярность может варьироваться у одного слова в зависимости от мнения автора и глагола который он выбрал для передачи информации. Таким образом тональные фреймы должны должны решать эту проблему посредством запоминания и извлечения таких связей из текста. В качестве основы для этой работы мы выбрали разработанный Лукашевич [1] тональный фреймворк.

Чтобы извлечь связи между объектами речи мы должны извлечь сами объекты речи. Первый вариант - это ручное маркирование, довольно энергозатратный и долгий метод, однако это компенсируется его точностью. Второй вариант - использование семантического парсера (DeepPavlov, UDPipe, Turku) [19], который может выделять объект и субъект самостоятельно и быстро, однако с более низкой точностью относительно ручного маркирования.

В данной работе будет предпринята попытка автоматического расширения словаря тональных фреймов с использованием современных методов машинного обучения, принимая во внимание все аспекты проблемы.

.

# **2.Ход работы**

*2.1. Данные*

В качестве основного материала работы был взят корпус твитов за авторством Рубцовой и словарь тональных фреймов Лукашевич (6707 глаголов). Для проверки гипотезы были взят корпус негативных твитов (111 923 вхождения) и положительных твитов (114 991 вхождений). В качестве предсказываемой величины было выбрано значение изменения полярности у объекта и субъекта глагола, проставленное экспертами. Эксперты оценивали каждый из глаголов и присваивали участникам ситуации значение: либо -1, либо 0, либо 1. Работа проводилась с двумя типами предсказываемой величины: как совместный класс, где величины полярностей складывались (классы -2, -1, 0, 1, 2) и убирались примеры, где обе полярности были равны нулю, и как два раздельных класса, где классификатору было необходимо рассчитать полярность сначала одного аргумента, а потом второго. Для проверки гипотезы необходимо было предобработать текст, с целью извлечь информацию о глаголах из твитов.

## *2.2.Базовая проверка метода*

Для базовой проверки метода, было решено использовать более простой набор утилит. В качестве токенизатора была использована функция из бибилиотеки “nltk”, а в качестве лемматизатора - библиотека “pymorphy2”.

NLTK [15] - это платформа основанная на Python для работы с текстовыми данными. В этой работе был использован токенизатор встроенный в эту библиотеку. Он основан на двух методах: TreebankWordTokenizer - метод использующий регулярные выражения для разбиения на токены, и PunktSentenceTokenizer - метод распознающий аббревиатуры и сокращения, для более точной токенизации текста.

Вторая библиотека, которая была использована в этом подходе - это Pymorphy [16], морфологический анализатор, основанный на данных собранных в словаре OpenCorpora. Этот модуль с помощью парадигм восстанавливает первоначальное слово и обозначает все категориальные данные.

С помощью этого подхода в каждом подкорпусе, негативном и позитивном, были выделены уникальные глаголы и метаинформация о них: сколько раз его формы встречались в текстах определённой полярности и сколько раз он встречался в определённом роде, числе, роде, времени и лице. Чтобы избежать шумов был установлен порог на 50 вхождений, если глагол встречался в негативных и позитивных твитах, не больше 50 раз, то такие данные отбрасывались, как неинформативные.

Далее было найдено пересечение получившегося списка глаголов из Твиттера со списком фреймовых глаголов. В качестве класса было решено взять изменение полярностей “а0” и “а1” под влиянием предиката. Экспертами была дана оценка изменения полярности, где “-1” означало, что аргумент получает негативную коннотацию из-за предиката, “0” означало, что аргумент не меняет своей полярности, и “1” означало, что аргумент становится положительным из-за предиката. *Рассказать про мультилейбел и однокласс (как собирал…)*

Как можно увидеть в Таблице 1, было выделено 13487 глаголов в негативном датасете и 12627 глаголов в позитивном датасете. Для проверки валидности взятого датасета необходимо было проверить насколько сильно материал твитов покрывает собранный словарь тональных фреймов. В результате проверки оказалось, что пересечение глаголов из положительного датасета с глаголами из фреймов составляет 27% от общего числа глаголов в фрейме, а пересечение глаголов из негативного датасета с глаголами из фрейма составляет 25%. Таким образом получается, что данные довольно равномерные и неплохо покрывают рабочий материал. Из-за изначальной неравномерности датасетов, положительных твитов на 3-4% больше, есть небольшой перекос в сторону положительных глаголов.

Таблица 1. Пересечение глаголов из твитов с глаголами из сентифреймов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Кол-во твитов | Кол-во глаголов | Пересечение с фреймом | Уникальных глаголов |
| Положительные | 114911 | 13487 | 27% (1799) | 370 |
| Негативные | 111923 | 12627 | 25% (1676) | 247 |

Из полученных данных был собран датасет, в котором каждой лемме глагола имеющейся в твитах соответствовала мета информация извлечённная из датасетов: встречаемость в позитивных и негативных подкорпусах, частота появления глагола в определённом лице, времени, числе и роде. Таким образом о глаголе была собрана наиболее полная информация, которую можно было извлечь данным методом.

В результате выполненной предобработки было создано два датасета: тренировочный (374 глагола) и тестовый(563 глагола). Все значения были переведены в проценты, дабы избежать ошибок и неправильного определения классов.

### *2.1.1 Мультиклассовая классификация*

Для проверки совмещённых данных было выбрано три различных классификатора, с помощью которых будет проводится анализ полученных данных: Логистическая регрессия, K-Neighbors, MLPClassifier.

Первый метод для предсказания класса в этой работе - логистическая регрессия [17]. Это модель предсказывающая вероятность принадлежности события к одному из классов. Также в работе будет рассмотрена работа классификатора по ближайшим соседям. Главный принцип этого метода, присваивание элементу класс, который наиболее распространён среди соседей этого элемента. Последний метод, который будет использован в этой работе - многослойный перцептрон. Это нейросетевой метод основывающийся на распределении весов между параметрами и получении предсказания для определённого события.

Тренировочный датасет был разбит на тренировочную и валидационную выборку, чтобы проверить работу классификаторов. Также датасет был дополнен автоматически сгенерированными примерами, для балансировки классов. Чтобы проверить предположение, какие из глагольных категорий могут влиять и указывать на направленность аргументов, было решено проверить как различные категории глаголов влияют на инструменты регрессии.

Логистическая регрессия показала неутешающие результаты на полученных данных (Таблица 2).

Таблица 2. Точность логистической регрессии без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| F1-macro | 0.40 | 0.40 | 0.34 | **0.44** | 0.36 |
| Accuracy | 0.43 | 0.44 | 0.35 | **0.46** | 0.38 |

Если посмотреть на распределение (Таблица 3) предсказанных классов в лучшем по точности варианте, то окажется, что регрессия не предсказывает классы -2, 0, 2.

Таблица 3. Распределение по классам в наиболее точном варианте логистической регрессии

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2 | -1 | 0 | 1 | 2 |
| Процент верных | 95 | 10 | 57 | 38 | 55 |

Следующим способом был метод К-ближайших соседей. Он показал результаты намного лучше чем в случае с логистической регрессией.(Таблица 4)

Таблица 4. Точность метода К-ближайших соседей без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| F1-macro | 0.71 | 0.72 | 0.72 | **0.73** | 0.69 |
| Accuracy | 0.71 | 0.72 | 0.72 | **0.73** | 0.69 |

Наилучшая точность (accuracy = 0.75, f1-macro = 0.75) отмечена при отсутствии показателей времени и числа (Таблица 5). С этими показателями классификатор довольно точно предсказывал классы 2, 0, -2.

Таблица 5. Распределение по классам в наиболее точном варианте метода К-ближайших соседей

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2 | -1 | 0 | 1 | 2 |
| Процент верных | 100 | 40 | 100 | 50 | 83 |

Следующий метод которым была проведена классификация стал многослойный перцептрон. Его предсказания оказались немного хуже чем метод К-ближайших соседей, однако, лучший результат был достигнут при наличии всех аргументов.

Таблица 6. Точность многослойного перцептрона без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| F1-macro | **0.68** | 0.66 | 0.65 | 0.66 | 0.65 |
| Accuracy | **0.67** | 0.66 | 0.66 | 0.67 | 0.65 |

Наилучший результат (Таблица 7) перцептрон показал при наличии всех данных

Таблица 7. Распределение по классам в наиболее точном варианте многослойного перцептрона

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2 | -1 | 0 | 1 | 2 |
| Процент верных | 100 | 29 | 93 | 46 | 85 |

В результате этого эксперимента мы увидели, что классификаторам оказалось сложнее предсказывать классы на стыке, где только один аргумент оказался направленным в какую-то из сторон. Также посмотрев на результаты можно предположить, что для данной задачи время оказалось скорее ненужной категорией глагола, так как с ней точность всегда ухудшалась, либо не менялась. Исходя из полученных результатов можно предположить, что если разбить задачу классификации на классификацию двух связных классов, то точность может повыситься для отдельно взятых случаев.

### *2.1.2 Политематическая классификация*

В случае раздельного маркирования обхекта и субъекта пополнять данные не пришлось из-за равномерного распределения трёх маркеров среди классов. Для политематической классификации был использован метода MultiOutputClassifier, который позволяет проводить обычные методы классификации на политематических данных.

После проведения экспериментов оказалось, что раздельное маркирование полярностей объекта и субъекта повысило точность на довольно значимую величину (Таблица 8)

Таблица 8. Точность методов политематической классификаци без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Accuracy of classification*** | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| К-ближайших соседей | **0.853** | 0.849 | 0.847 | 0.845 | 0.85 |
| Логистическая регрессия | 0.827 | 0.842 | 0.847 | **0.849** | 0.847 |
| Многослойный перцептрон | 0.752 | **0.774** | 0.693 | 0.681 | 0.676 |

*Таблица 8*

Таким образом на политематической классификации лучше всего себя показал метод К-ближайших соседей, который основывался на всех категориях глагола.

В результате проведённых исследований на примере базового метода, можно прийти к выводу, что совмещённая оценка субъекта и объекта как показатель менее информативна и менее предсказуема. Также можно заметить, что категория времени во многих случаях выступала помехой для классификатора, следовательно является менее информативной на получившимся датасете.

Таким образом уже с помощью базового метода можно относительно неплохо предсказывать полярности получившегося датасета.

### *2.1.3. Исследование предсказаний на неразмеченных данных*

При оценке получившейся классификации, можно заметить большое количество нейтральных глаголов, для которых было предсказано какое-то из значений для аргументов. Например, случаи с“зайти”, “встать”, “звонить” получили негативную оценку для объекта, а случаи с“краситься” и рулить получили положительную оценку для объекта. Несмотря на большое количество такого шума, можно заметить, что много глаголов у которых есть явное влияние на свои аргументы были размечены правильно. Так классификатор выделил слово в случае со словом “улыбнуть” отметил положительное влияние глагола на оценку как субъекта так и объекта.

(1) “*Как ты ее от других корейцев отличаешь..? - Да у нее глаза большие” Улыбнуло :)*

В случае с глаголом “достаться” классификатор выделил положительное изменение субъекта, однако возможная негативная коннотация в отношении объекта не была выделена

(2) *Сёдня жeрeбьёвka на ЧМ-2014.Хоть бы России шлak в группe достался)*

У глагола устать был выделен негативный оттенок у объекта.

(3) *Устала так скучно жить хоть бы в школу пошла нет \*\*\*\*\* сиди дома)*

Интересный случай произошёл с глаголом попасть. После работы классификатор оценил его объект как негативный. Скорее всего тут повлияла специфика лексики Твиттера, приняв более разговорное значение, со смыслом попасть в неприятную ситуацию.

(4) *@Jyliaalex смотри там случайно в аспирантуру не попади, наука никого до добра не доводила)*

Подводя итоги исследования на примере этого метода, можно сказать, что выбранный способ оказался довольно точным, хотя и требует улучшений. Однако оценивая разнообразие предсказанных классов для субъекта глагола, оказалось, что ни один из субъектов не был отмечен как отрицательный

## *2.2.Основной метод*

Главным методом в этой части работы стал семантический парсер и анализатор - DeepPavlov [18]. DeepPavlov - это модель искусственного интеллекта, предназначенная для анализа текста. Главным преимуществом данной модели является гибкость к заадчам и многопрофильность.

Результатом работы этого алгоритма стал обработанные тексты из Твиттера в формате CoNLL-U. Из текста были выделены глаголы и все возможные связи между частями предложений. Была также проведена проверка на валидность собранных данных (Таблица 9)

Таблица 9. Пересечение глаголов из твитов с глаголами из сентифреймов(DeepPavlov ver.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Кол-во твитов | Кол-во глаголов | Пересечение с фреймом | Уникальных глаголов |
| Положительные | 114911 | 14956 | 28% (1858) | 372 |
| Негативные | 111923 | 14335 | 26% (1740) | 254 |

Можно заметить, что DeepPavlov чуть лучше обработал текст, найдя чуть больше глаголов, чем pymorphy. Также, выросло количество новых глаголов после обработки информации и избавления от шумов, тренировочный датасет вырос до 379 вхождений, тестовый датасет без размеченных классов до 602 вхождений.

### *2.2.1 Политематическая классификация на примере DeepPavlov*

В качестве нового параметра в классификации появился залог, извлечённый из категорий глагола. Результаты получившиеся по итогам работы трёх классификаторов (Таблица 10) оказались чуть лучше, однако, назвать это статистически значимым нельзя.

Таблица 10. Точность методов политематической классификаци без различных категорий

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Accuracy of classification*** | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица | Без залога |
| К-ближайших соседей | 0.832 | 0.827 | 0.834 | **0.842** | 0.83 | 0.833 |
| Логистическая регрессия | 0.792 | 0.791 | 0.803 | **0.804** | 0.776 | 0.812 |
| Многослойный перцептрон | **0.753** | 0.673 | 0.729 | 0.673 | 0.702 | 0.708 |

Однако, лучшим результаты был достигнут при выбрасывании из датасета показателя времени и рода и использовании метод К-ближайших соседей, Accuracy достигла 0.854.

Можно заметить, что улучшение точности метода выделения категорий глагола, не дала значительного прироста в точности, существующий прирост скорее можно считать статистической погрешностью, которая может меняться в зависимости от разбиения выборки на тестовую и валидационную, однако результаты полученные в ходе предсказаний в совмещении с DeepPavlov дают интересный материал для исследования.

### *2.2.2. Исследование предсказаний на неразмеченных данных*

В предсказаниях на основе семантического и морфологического анализатора Deep Pavlov появилось больше разнообразия в классах. Вследствие этого несмотря на не большой прирост в точности на примере валидационной выборки, результаты на тестовой получились более правдоподобные.

Интересный пример со словом “ожидать”, который классификатор разметил как разнонаправленный, субъекту, “а0”, был присвоен положительный класс, а объекту, “а1”, был присвоен негативный. В прошлом исследовании классификатор обозначил лишь негативный оттенок у объекта, но положительный у субъекта не выдал.

У глаголов “послать” и “слать” в результате классификации у объекта выделена негативная оценка. Тут скорее всего как и в случае с “попасть” большое влияние имела разговорная направленность твиттера, будучи источником разговорной короткой речи, множество слов принимают значения обыденные для устной речи.

(5) *ПЕРЕПОСТ!!!!!!!!!!!!!! ЖАЛЬ, ЧТО НА УКРАИНУ НЕ ПОСЛАЛИ)))* [*http://t.co/W0uqzLaKqA*](http://t.co/W0uqzLaKqA)

Однако на более точную оценку неразмеченных глаголов, классификация нейтральных глаголов тоже оставляет желать лучшего: “быть”, “мочь”, “смотреть” размечены так, что объект глагола получает положительную оценку.

Глагол “разбудить” разметился как явно негативный для субъекта глагола. А в случае “кричать” негативным оказался объект, как виновник некоего проишествия.

# **3. Результаты**

# В ходе исследования была проверена гипотеза о возможности автоматического извлечения тональных фреймов. Работа проводилась над определением изменения полярностей у аргументов предиката. Лучшим результатом на валидационной выборке с точностью 85,4% стал метод К-ближайших соседей на основе данных полученных при работе анализатора DeepPavlov. После проведённого анализа результатов предсказаний на тестовой выборке, можно сказать, что классификатор работает с приличной точностью, но недостаточной, чтобы на неё опираться в следующих работах. Приведённый метод нуждается в дополнении и улучшении.

Твиттер показал свою состоятельность как источник для данных, выборки получились информативными, точность оказалось неплохой для автоматизированного метода работы. Субъективная и насыщенная эмоциями лексика оказалась неплохим материалом для анализа и поиска связей между данными извлечёнными из предложений и полярностью аргументов глагола.

# **4. Заключение**

Работа проведённая в этом исследовании далека от идеала, не было проведено полного анализа текста, чтобы найти определённые паттерны связи изменения полярности у аргументов предиката, также методы классификации тоже можно улучшить. Однако, даже с помощью такого пайплайна можно добиться неплохих результатов в поиске паттернов.

Одной из задач по улучшению этого метода можно отметить очистку данных от нейтральных глаголов, которые будут лишь вызывать шумы при классификации неразмеченных данных.

Полученную классификацию нельзя назвать в полной степени тональным фреймом, так как покрывает это распределение лишь часть информации доступной в датафрейме, но полученную информацию уже можно считать подтверждением гипотезы, что есть возможность без помощи большого количества экспертов автоматически извлекать тональные фреймы из текста и использовать их при анализе настроения текстов.

# **5. Список литературы**

1. Карнаухова В.А., Лукашевич Н.В., Принципы создания системы оценочных фреймов для русского языка - 2018
2. Jurafsky D. Speech & language processing. – Pearson Education India, 2000.
3. Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques //Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. – Association for Computational Linguistics, 2002. – С. 79-86.
4. Snyder B., Barzilay R. Multiple aspect ranking using the good grief algorithm //Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. – 2007. – С. 300-307.
5. Рубцова Ю.В. Методы автоматического извлечения терминов в динамически обновляемых коллекциях для построения словаря эмоциональной лексики на основе микроблоговой платформы Twitter // Доклады Томского государственного университета систем управления и радиоэлектроники. 2014, № 3 (33). –С.140-144.
6. Stone P. J., Dunphy D. C., Smith M. S. The general inquirer: A computer approach to content analysis. – 1966.
7. Thongtan T., Phienthrakul T. Sentiment Classification using Document Embeddings trained with Cosine Similarity //Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. – 2019. – С. 407-414.
8. Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //arXiv preprint arXiv:1910.10683. – 2019.
9. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.
10. Pang B., Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts //Proceedings of the 42nd annual meeting on Association for Computational Linguistics. – Association for Computational Linguistics, 2004. – С. 271.
11. Liu B. et al. Sentiment analysis and subjectivity //Handbook of natural language processing. – 2010. – Т. 2. – №. 2010. – С. 627-666.
12. Rashkin H., Singh S., Choi Y. Connotation frames: A data-driven investigation //arXiv preprint arXiv:1506.02739. – 2015.
13. Xie Q. et al. Unsupervised data augmentation for consistency training. – 2019.
14. Mohammad S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2018. – С. 174-184.
15. Loper E., Bird S. NLTK: the natural language toolkit //arXiv preprint cs/0205028. – 2002.
16. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 320-332.
17. Kleinbaum D. G. et al. Logistic regression. – New York : Springer-Verlag, 2002.
18. Burtsev M. et al. Deeppavlov: Open-source library for dialogue systems //Proceedings of ACL 2018, System Demonstrations. – 2018. – С. 122-127.
19. Straka M., Hajic J., Straková J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing //Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). – 2016. – С. 4290-4297.

# **6. Приложение**

Ссылка на код и материалы работы: <https://github.com/NanairoNika/GCW>

Таблицы из текста работы:

Таблица 1. Пересечение глаголов из твитов с глаголами из сентифреймов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Положительные | 114911 | 13487 | 27% (1799) | 370 |
| Негативные | 111923 | 12627 | 25% (1676) | 247 |

Таблица 2. Точность логистической регрессии без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| F1-macro | 0.40 | 0.40 | 0.34 | **0.44** | 0.36 |
| Accuracy | 0.43 | 0.44 | 0.35 | **0.46** | 0.38 |

Таблица 3. Распределение по классам в наиболее точном варианте логистической регрессии

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2 | -1 | 0 | 1 | 2 |
| Процент верных | 95 | 10 | 57 | 38 | 55 |

Таблица 4. Точность метода К-ближайших соседей без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| F1-macro | 0.71 | 0.72 | 0.72 | **0.73** | 0.69 |
| Accuracy | 0.71 | 0.72 | 0.72 | **0.73** | 0.69 |

Таблица 5. Распределение по классам в наиболее точном варианте метода К-ближайших соседей

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2 | -1 | 0 | 1 | 2 |
| Процент верных | 100 | 40 | 100 | 50 | 83 |

Таблица 6. Точность многослойного перцептрона без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| F1-macro | **0.68** | 0.66 | 0.65 | 0.66 | 0.65 |
| Accuracy | **0.67** | 0.66 | 0.66 | 0.67 | 0.65 |

Таблица 7. Распределение по классам в наиболее точном варианте многослойного перцептрона

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | -2 | -1 | 0 | 1 | 2 |
| Процент верных | 100 | 29 | 93 | 46 | 85 |

Таблица 8. Точность методов политематической классификаци без различных категорий

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Accuracy of classification*** | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица |
| К-ближайших соседей | **0.86** | 0.85 | 0.85 | 0.85 | 0.85 |
| Логистическая регрессия | 0.83 | 0.84 | 0.85 | 0.85 | 0.85 |
| Многослойный перцептрон | 0.75 | 0.77 | 0.69 | 0.68 | 0.68 |

Таблица 9. Пересечение глаголов из твитов с глаголами из сентифреймов(DeepPavlov ver.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Кол-во твитов | Кол-во глаголов | Пересечение с фреймом | Уникальных глаголов |
| Положительные | 114911 | 14956 | 28% (1858) | 372 |
| Негативные | 111923 | 14335 | 26% (1740) | 254 |

Таблица 10. Точность методов политематической классификаци без различных категорий

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Accuracy of classification*** | Все данные присутствуют | Без числа | Без рода | Без времени | Без лица | Без залога |
| К-ближайших соседей | 0.832 | 0.827 | 0.834 | **0.842** | 0.83 | 0.833 |
| Логистическая регрессия | 0.792 | 0.791 | 0.803 | **0.804** | 0.776 | 0.812 |
| Многослойный перцептрон | **0.753** | 0.673 | 0.729 | 0.673 | 0.702 | 0.708 |