Классификация несловарных слов в русском языке: автоматическая обработка данных и сравнение эффективности подходов

Алексей Доркин

Владислава Смирнова

Антон Вахранёв

Исследовательская работа посвящена изучению несловарных слов и необычных морфологических структур, которые можно встретить в современном русскоязычном сегменте интернета, и классификации этих форм. Авторы статьи работали над датасетом, состоящим из 47 миллионов словоформ, собранных из ГИКРЯ. Данные включают слова, намеренно написанные в искаженном виде (эрративы), экспрессивные формы и заимствования из английского языка – эти группы представляют для авторов особый интерес. Для каждого класса слов были выявлены характерные признаки, использованные для обучения классификаторов. Задачи бинарной классификации по 3 целевым классам по отдельности были выполнены с помощью готовых алгоритмов. После проведения анализа результатов первичной классификации лучшие модели были использованы для создания баз искажений, англицизмов и экспрессивных форм.

Ключевые слова: продуктивные морфологические классы, заимствования, искажения, экспрессивная лексика

Classification of non-dictionary words in Russian: comparison of the effectiveness of different approaches of data processing

Key words: productive morphological classes, loanwords, deforming words, expressive words

Оглавление

1. Введение

Изучение лексики, возникшей и использующейся именно в интернете, – в целом довольно молодое направление в науке. Так, самые ранние из найденных нами исследований на подобную тематику датируются 2008 годом[1]. Предметом подобных найденных нами исследований 2010-х годов был “олбанский язык” (сленг русскоязычного сегмента интернета, был распространен в 2000-х годах, обладает особой антинормативной орфографией и рядом собственных устойчивых выражений). Среди более новых работ встречаются как те, в которых преимущественно описан сленг интернета (примером таковой является книга “Словарь языка Интернета.RU”[2]), так и те, что изучают более широкий спектр слов, которые можно встретить в интернете, или морфологические процессы интернет-лексики (пример таковой – “Современный русский язык в интернете”[3]). Наша работа состоит в изучении не только сленга, но и другой лексики.

Тем не менее, наша работа имеет определенные ограничения: она сконцентрирована на трех категориях слов, представляющих для нас наибольший интерес. К ним относятся пришедшие из английского языка заимствования, экспрессивная лексика и эрративы.

Одна из причин, по которым мы решили сосредоточиться именно на этих категориях слов, заключается в больших возможностях для образования новых слов в данных классах. Мы предполагаем, что слова этих категорий высоко продуктивны (то есть что на основе этих слов и их морфем легко образовывать новые слова). Также мы предполагаем, что среди неологизмов и окказионализмов, которые можно встретить в интернете, большой процент слов попадет именно в наши целевые классы.

Многие существующие по такому направлению работы основываются на нормативном русском языке, то есть в них используются данные из словарей, НКРЯ и других похожих корпусов. Примеры таких работ: “Анализ функционирования экономических заимствований в современных российских СМИ”[4] (статья посвящена краткому анализу функционирования англицизмов в российских СМИ на примере экономических терминов, анализ результатов подкреплён исследованиями на основе НКРЯ), “Намеренное нарушение языковой нормы: исследование на основе языковых

[1] Карасева, А.И. Роль и функции эрратива в интернет-сленге // Мониторинг общественного мнения: экономические и социальные перемены. 2008. №2(86). С. 129-140.
[2] Словарь языка Интернета.RU / Кронгауз М. А., Пиперски А. Ч., Сомин А. А., Черненко Ю. А., Мерзлякова В. Н., Литвин Е. А., Под ред. Кронгауза М. А. М.: АСТ-Пресс Книга, 2016.
[3] Современный русский язык в интернете / Под ред. Я. Э. Ахапкина, Е. В. Рахилина. М.: Языки славянской культуры, 2014.
[4] Косенко Е.И. Анализ функционирования экономических заимствований в современных российских СМИ // Вестн. Сев. (Арктич.) федер. ун-та. Сер.: Гуманит. и соц. науки. 2017. № 2. С. 107-113.

корпусов”[5] (исследование проводилось на материалах НКРЯ и BNC) и “Механизмы экспрессивности в языке”[6]. Мы же анализируем слова из корпуса ненормативных текстов, взятых из разных интернет-источников и добавленных в Генеральный Интернет-Корпус Русского Языка (ГИКРЯ[7]).

Мы изучаем отдельные слова, собранные из ГИКРЯ Ириной Кротовой и Василисой Андриянец для работы “Antidictionary: database of out-of-dictionary words from the Russian Internet”[8]. В этих данных слова представлены без контекста. Размер исходного датасета составлял 47 258 521 словоформ. После предобработки, о которой подробно написано в пункте *3. Предобработка данных*, в нем осталось 6 647 276 уникальных словоформ.

Основная идея нашей работы состоит в исследовании и бинарной классификации несловарных слов, которые можно встретить в современном русскоязычном интернете и которые могут быть неправильно распознаны или не распознаны вовсе крупными морфологическими парсерами, такими, как, например, *MyStem*[9]. Поэтому цели нашего исследования – обучить разные алгоритмы бинарной классификации распознавать интересующие нас классы слов в несбалансированном датасете, собрать показатели работы выбранных нами моделей, отобрать из них наиболее эффективные по оценке метрик (*примечание: описание выбранных нами метрик находится в части 4. Выбор инструментов и методов*) и с их помощью создать базу несловарных форм, состоящую из англицизмов, экспрессивной лексики и слов с искажениями. Предполагается, что в получившейся в итоге базе можно будет найти новые, не найденные нами подкатегории трех целевых классов.

2. Обзор литературы

По всем исследуемым в этой работе направлениям небольшое количество научной литературы, что в значительной мере подчёркивает актуальность данного исследования.

Среди работ по изучению англицизмов наиболее близким к нашей работе было исследование Феногеновой, Карпова, Казорина и Лебедева “Сравнительный анализ

---

[5] Вахранев А.Ю. Намеренное нарушение языковой нормы: исследование на основе языковых корпусов // Пространство научных интересов:иностранные языки и межкультурная коммуникация – современные векторы развития и перспективы: сборник статей по результатам IV научной межвузовской конференции молодых ученых 11.04.2019 г. (ДИЯ НИУ ВШЭ). Отв. редактор Е.Г. Кошкина. М.: ООО «Буки Веди», 2019. С. 32-40.

[6] Зализняк А. А. Механизмы экспрессивности в языке // Ю.Д. Апресян и др. (ред.). Смыслы, тексты и другие захватывающие сюжеты: Сб. ст. в честь 80-летия Игоря Александровича Мельчука. — М.: ЯСК, 2012. — С. 650–664.

[7] О проекте // Генеральный Интернет-Корпус Русского Языка URL: http://www.webcorpora.ru (дата обращения: 14.04.2020).

[8] Antidictionary: database of out-of-dictionary words from the Russian Internet // AINL 2018 – Agenda. Poster and demo session. Posters URL: https://ainlconf.ru/2018/agenda#program-demo-and-posters (дата обращения: 26.06.2020).

[9] MyStem // Яндекс URL: https://yandex.ru/dev/mystem/ (дата обращения: 14.04.2020).

распределения англицизмов в русских текстах социальных медиа”[10]. Эта статья продолжает предыдущую работу авторов по обнаружению англицизмов, предлагая улучшенный метод поиска заимствованных слов с помощью современных методов машинного перевода. Проведённое авторами сравнительное исследование в Живом Журнале, Вконтакте, Хабрахабре и Твиттере показывает, что разные социальные, гендерные и даже возрастные группы имеют одинаковую долю англицизмов в речи.

Работ по извлечению экспрессивных форм также обнаружено не было, однако в таких работах, как “Извлечение именованных сущностей с помощью Википедии”[11] Д.И. Астаховой, можно найти пример того, как для извлечения используются данные о суффиксах и префиксах слов. В данной работе решается задача применения в системе извлечения именованных сущностей для русского языка информации, получаемой из Википедии. Автором был предложен и реализован способ извлечения информации из Википедии, а также предложены признаки, использующие извлечённую информацию, и реализовано извлечение таких дополнительных признаков в системе извлечения именованных сущностей.

Работа Бориса Валерьевича Орехова “Суперминимум и нанодержава: префиксоиды в языке интернета”[12] также построена на извлечении слов с конкретными префиксами (в том числе — с экспрессивными: “*супер-”, “мега-”, “гипер-”*) из датасета, собранного из русскоязычных блогов и текстов интернет-СМИ Lenta.ru.

Если же говорить о работе моделей в целом, то здесь можно обратить внимание на работу “Enriching Word Vectors with Subword Information” (Bojanowski, Grave, Joulin, Mikolov), посвящённую обогащению словарных векторов информацией, которая содержится в подсловах (т.е. морфологическая информация).

Проблемы, связанные с извлечением слов с орфографическими ошибками, возникают не только в научных работах, связанных с NLP, но и в профессиональных задачах. О наличии таких проблем свидетельствуют появляющиеся на профессиональных блоговых площадках материалы[13] на подобную тематику. Работ по извлечению намеренных искажений нами обнаружено не было.

3. Предобработка данных

Исходный датасет включает три значимых столбца:

---

[10] Феногенова А.С., Карпов И.А., Казорин В.И., Лебедев И.В. Сравнительный анализ распределения англицизмов в русских текстах социальных медиа // Сборник Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 1 — М.: Изд-во РГГУ, 2017. С. 65-74.

[11] Астахова Д.И. Извлечение именованных сущностей с использованием Википедии // Московский государственный университет имени М.В. Ломоносова. Факультет вычислительной математики и кибернетики. Кафедра системного программирования. Москва, 2015.

[12] Орехов Б.В. Суперминимум и нанодержава: префиксоиды в языке интернета // Современный русский язык в интернете / ред. Я. Э. Ахапкина, Е. В. Рахилина. — М.: Языки славянской культуры, 2014. С. 281-290.

[13] Задача: извлечь ключевые выражения из текста на русском языке. NLP на Python // Habr URL: https://habr.com/ru/post/468141/ (дата обращения: 20.03.2020).

1. слово;
2. предположительные морфологические характеристики слова в виде тега;
3. предположительная лемма, которую можно получить с помощью MyStem, обработав слова из первой колонки.

Рис. 1. Случайная выборка из основного датасета

	number	word	tag	lemma
1493483	8402118	Олливандером	Ncmsin	Олливандером
976628	8028190	Колоновидные	Afpmpnf	Колоновидные
903574	2634985	Карабос	Ncmsny	Карабос
4196075	7948071	качественная	Afpfsnf	качественная
4692906	9087985	мимимикрируют	Vmip3p-a-e	мимимикрируют
6368019	3187575	рассылка	Afpfsns	рассылка
6290785	3961090	ражиссер	Ncmsny	ражиссер
6243645	4666518	пугинский	Afpmsaf	пугинский
6350659	9783476	раскепостившихся	Afpmpgf	раскепостившихся
2146936	2464704	Танченг	Ncmsan	Танченг

Нами использовался только столбец “word” с исходными словоформами, однако остальные столбцы были сохранены на тот случай, если в дальнейшем в них возникнет необходимость. Информация из колонки “tag” не учитывалась, поскольку одной из задач данной работы было изучение характерных для наших целевых классов сочетаний символов (символьных n-грамм), а не грамматические и морфологические признаки.

Этапы предобработки, которые мы использовали в работе с данными, следующие.

1) Удаление идентичных строк.

Этот этап необходим для сокращения размера датасета и, как следствие, для сокращения объемов памяти, необходимых для работы с данными.

2) Преобразование строк к декомпозированной форме Unicode-символов

Данный этап необходим для корректной сортировки слов с буквой “ё”.

3) Сортировка строк в алфавитном порядке по столбцу “word” (слову).

Этот этап нужен для удобства оценивания количества оставшихся похожих строк, которые могут отличаться друг от друга 1-3 символами.

4) Преобразование строк обратно в композированную форму.

Шаг необходим для дальнейшей корректной работы над данными.

5) Поиск слов, содержащих три и более повторения одного символа подряд и сжатие повторяющихся букв в одну в таких словах, как “ЁЁЁлки”, “АААААОБОЖАЮ”, “учусссь”.

Необходимость в сокращении некоторых символов обусловлена наличием дублирующихся слов, отличающихся друг от друга только 1-3 буквами (например, “ЁЁЁлки” – ”ёлки”) или же многократным повторением одних и тех же букв (“ммммаааааммммаааа” – “мама”). Для более точной оценки интернет-лексики мы приняли решение привести такие слова к одинаковому виду и убрать полные дубликаты, появившиеся после сокращения символов. Далее в статье мы объясняем, что после обработки некоторые слова в данных все же содержат размноженные буквы. Мы приняли решение не сокращать все размноженные буквы во всех словах, потому что слишком большое сокращение могло исказить такие слова, как, например, “фароссских”. Если сократить 3 буквы “с” в этом слове в одну, то слово будет написано с ошибкой. Такие ошибки могли повлиять на наше исследование и работу моделей. При этом наличие в слове размноженных букв не являлось в рамках нашего исследования маркером искажения или экспрессивности слова.

6) Если у слова сокращалось число букв, то оно записывалось в отдельный столбец, если же слово не менялось, то оно записывалось в отдельный столбец в идентичном виде.

Таким образом у нас получился отдельный столбец с обработанными словами.

7) Удаление дубликатов, а также строк, в которых слова в новом столбце короче трёх букв.

Например, слова, состоящие из одной буквы (“ЁЁЁ”, “ААА”), слова, похожие на аббревиатуры слов (“АХШ”, “яхз”), наборы символов (“АХа”, “ГгГ”) не представляют для данной работы особого интереса.

8) Удаление слов, которые обнаруживаются в словаре Зализняка[14].

В рамках нашей работы не представляют особого интереса словарные формы, такие, как, например, “графского”, “куролесившие”, “матрёшкам”.

9) Приведение всех обработанных слов к нижнему регистру.

[14] Словарь Зализняка oDict.ru // Открытый грамматический словарь русского языка URL: http://odict.ru (дата обращения: 10.09.2019).

Для нас изменение регистра слов не является предметом исследования, но мы допускаем, что дальнейшие работы могут быть связаны с изучением изменения регистра при написании слов в интернете.

После всех этапов обработки данных мы получили 6 647 276 уникальных словоформ, которые использовали в дальнейшей работе.

Для создания тренировочных множеств нецелевых классов мы использовали словарь Зализняка. Причина, по которой мы выбрали именно его, заключается в том, что он состоит не только из лемм, но и из словоформ не в начальной форме. Поскольку в наших данных присутствуют слова, для которых сложно определить леммы, а также слова в разных грамматических формах, мы выбрали именно такой вариант словаря[15]. В качестве этапов предобработки словаря мы убрали все, что не являлось основным содержимым словаря (словоформами), и привели все формы к нижнему регистру, так как в наших обработанных данных все слова тоже были приведены к нижнему регистру.

В качестве обучающего множества заимствований мы выбрали словарь англицизмов русского языка Дьякова[16], потому что он содержит 16134 слов, что больше, чем то количество слов, которое мы могли бы собрать вручную из датасета. Также для нас важно, что этот словарь полностью состоит из заимствований из английского языка, поскольку в рамках данной работы мы сосредоточились именно на таких заимствованиях. Предобработка словаря Дьякова заключалась в удалении того, что не являлось основным содержимым словаря (словами), удалении дубликатов и приведении строк к нижнему регистру.

Удаление дубликатов и приведение слов к нижнему регистру – это также те этапы, которые мы использовали в предобработке слов для обучающих и тестовых множеств эрративов и экспрессивных слов. Подробнее дополнительные источники этих данных описаны в соответствующих частях (*6. Работа над словами с искажениями* и *7. Работа над экспрессивными формами)*. Основными источниками тренировочных данных являлся исследуемый нами датасет, из которого мы вручную извлекали подходящие слова.

4. Выбор инструментов и методов

В рамках этой работы мы приняли решение выполнить бинарную классификацию и работать с тремя целевыми категориями слов по отдельности. Для получения опорной (baseline) метрики классификации была использована логистическая регрессия. Для того, чтобы превысить значения baseline классификатора, помимо регрессии мы попробовали использовать метод опорных векторов и

[15] Словарь Зализняка oDict.ru // Открытый грамматический словарь русского языка URL: http://odict.ru (дата обращения: 10.09.2019).
[16] Словарь // Дьяков А.И. Словарь англицизмов русского языка URL: http://anglicismdictionary.ru/Slovar (дата обращения: 10.09.2019).

стохастический градиентный спуск (SGD), а также ансамблевую модель – случайный лес (Random Forest). Дополнительно в работе с заимствованиями мы попробовали обучить сверточную сеть (*примечание: эксперименты с нейронными сетями не улучшили метрики, о чем подробнее написано в части 5. Работа над заимствованиями, поэтому мы решили не продолжать работу с сетями*).

Для работы с классификаторами было необходимо получить векторные представления слов из датасета. Для решения данной задачи мы выбрали *fastText* с сайта RusVectōrēs[17], обученный на корпусе Araneum Russicum. Выбор именно этой версии *fastText* связан с тем, что Araneum Russicum – это интернет-корпус русского языка[18]. Соответственно, он в меньшей степени состоит из литературных и новостных источников и содержит не только словарные слова, но и интернет-лексику. Полагаем, что это является его преимуществом для наших задач, связанных с несловарными формами из интернета.

Выбор *fastText* также мотивирован тем, что, в отличие от других моделей, таких, как, например, *Word2Vec*, для получения эмбеддингов *fastText* использует не только дистрибутивные характеристики самого слова, но и подслова (subwords), в него входящие[19].

Нами было принято решение использовать предобученную модель также в связи с тем, что обучение своей модели *fastText* связано с высокими требованиями к вычислительным мощностям и объему обучающего корпуса.

Для оценки качества работы классификаторов была выбрана f-мера.

Формула f-меры

$$precision = Pr = \frac{tp}{tp+fp} \text{ – точность}$$

$$recall = R = \frac{tp}{tp+fn} \text{ – полнота}$$

$$F_2 = \frac{2Pr*R}{Pr+R} \text{ – } F\text{-мера}$$

,

где tp – true positive, fp – false positive, tn – true negative, fn – false negative.

Во-первых, это связано с тем, что она включает в себя сразу две метрики – точность и полноту. Во-вторых, она подходит для оценки работы модели на несбалансированных данных. Поскольку в нашем датасете словарных слов, а также форм, которые не относятся к нашим целевым классам, может быть намного больше, чем интересующих нас слов, можно назвать наши данные несбалансированными. К

[17] Статические модели // RusVectōrēs: семантические модели для русского языка URL: https://rusvectores.org/ru/models/ (дата обращения: 04.04.2020).
[18] Другие корпуса // Национальный корпус русского языка URL: http://www.ruscorpora.ru/new/corpora-other.html (дата обращения: 14.04.2020).
[19] Importance of character n-grams // Word representations. FastText Docs URL: https://fasttext.cc/docs/en/unsupervised-tutorial.html#importance-of-character-n-grams (дата обращения: 15.09.2019).

сожалению, точно оценить пропорции разных классов до начала исследования невозможно по причине того, что мы работаем с полностью не размеченными данными.

Для более точной оценки модели мы воспользовались работой волонтеров-аннотаторов. Мы предложили им оценить случайные выборки из трех баз размеченных слов. Аннотаторы приписывали слову число 1, если оно принадлежит к одному из целевых классов, и 0, если оно относится к другой категории. Оценка принадлежности слов к интересующим нас категориям проходила независимо, для каждого класса по отдельности. На основе размеченных слов мы посчитали коэффициент альфа Криппендорффа – статистическую меру оценки согласия между аннотаторами, которая показывает, насколько разметка моделью точна. Коэффициент принимает значение от 0 до 1.

Формула альфы Криппендорфа

$$\alpha = 1 - \frac{D_o}{D_e},$$

где D_o – наблюдаемое разногласие, D_e – ожидаемое разногласие (такое, которое можно получить случайно).

Для итоговой классификации данных была использована модель, использующая метод опорных векторов и стохастический градиентный спуск, поскольку она показала наилучший результат по f-мере для всех категорий слов. Значения метрик, получившихся в результате работы каждой модели, указаны в частях *5. Работа над заимствованиями*, *6. Работа над словами с искажениями* и *7. Работа над экспрессивными формами*.

Мы считаем критерием неуспешного анализа слов значение f-меры ниже 0.50, поскольку такой показатель будет означать неправильное и/или случайное извлечение данных моделью. А также альфа Криппендорфа ниже 0.5, что будет означать отсутствие согласия между аннотаторами в оценке работы модели. Предполагается, что чем более точно модель находит целевые слова, тем больше аннотаторы согласны в оценивании работы модели.

5. Работа над заимствованиями

Конечной целью нашей работы является получение из исходных данных наборов слов, отобранных по определенному признаку или признакам. Самой очевидной группой слов представляются заимствования из других языков (в рамках нашей работы, английского). Такая оценка обусловлена кажущейся простотой выделения из лексикона русского языка отдельной группы слов, имеющих иной источник (или источники) происхождения относительно остальных. Делается предположение, что такие слова имеют последовательности букв менее распространенные в “исконно русских” словах или даже “обрусевших” заимствованиях

(т.е. речь идет об относительно ранних заимствованиях, источником которых, как правило, является не английский язык), что может служить формальным признаком для их выделения.

Фактически речь в данном случае идет о словах из английского языка, написанных кириллицей по разным принципам. Некоторые слова записаны так, как они произносятся. Другие основаны на написании и последующем прочтении не по правилам английского языка. Некоторые примеры таких слов приведены в таблице ниже.

Таб. 1. Примеры заимствований

Пример
имиджмейкинг
формер
вертекса
старлингу
нойзгитар
офпати
ипад
фулашди
экшнам
глобалгивинг
скайфокс
даунлоад

Для получения из исходного набора данных слов, которые можно определить как заимствования, необходим классификатор. Поскольку данные не были размечены, возникла необходимость использовать отдельные наборы данных для классификации.

Для создания группы слов, не являющихся заимствованиями, был использован словарь Зализняка. Для являющихся – словарь Дьякова.

Стоит отметить, что словарь Дьякова по большей части состоит именно из слов, подобных приведенным в примере выше.

Объем словаря Зализняка значительно превосходит по объему словарь Дьякова, что приводит к сильной несбалансированности данных, используемых для обучения. Одним из решений такой проблемы является сэмплирование данных и изменение итогового соотношения классов. При обучении моделей использовалась случайным образом отобранная подвыборка из словаря Зализняка и целиком словарь Дьякова. Соотношение заимствований и не заимствований составило в результате 1 к 6. Такое соотношение было выбрано с той целью, чтобы сохранить несбалансированность реально имеющихся данных, но при этом не жертвовать качеством модели. Оценить реальное соотношение словарных и несловарных слов в живом языке представляется довольно затруднительным, а кроме такая оценка не является целью данной работы. По этой причине репрезентативность используемых в работе выборок не постулируется.

Для каждого слова было получено векторное представление с помощью предобученной модели fastText. Далее эти представления были использованы для классификатора, использующего логистическую регрессию. Если классификатор присваивал слову класс 1, то оно сохранялось в отдельный файл.

Также в рамках работы над данной задачей, было опробовано несколько классификаторов, как линейных, так и нелинейных, входящих в библиотеку *sklearn* для языка программирования Python: Random Forest, SGD Classifier, Multilayer Perceptron Classifier и некоторые другие. Изначально была опробована логистическая регрессия, f-мера которой составила 0.78. Оказалось, что лучший результат дает SGD Classifier, значение f-меры которого составило 0.85. Результат многослойного перцептрона варьировался от 0.84 до 0.85 при разных запусках (random seed не фиксировался), однако обучался он заметно дольше. Результат Random Forest классификатора составлял 0.49.

Также, в качестве эксперимента, была обучена посимвольная сверточная нейронная сеть с простой архитектурой, без использования предобученных эмбеддингов fastText. Полученная f-мера составляла от 0.5 до 0.52. Данный результат варьируется от запуска к запуску, поскольку при использовании нейронных сетей важную роль играет случайность при установке изначальных весов, поэтому фиксировать random seed не принято, что, в свою очередь, ведет к разбросу в получаемой оценке качества.

Чтобы оценить степень соответствия полученного классификатора нашим ожиданиям, слова, полученные методом описанным выше, была размечены независимыми аннотаторами.

На полученных данных была оценен коэффициент согласия аннотаторов – альфа Криппендорффа, которая составила 0.55. Полученное значение говорит, что между аннотаторами не был достигнут консенсус. Из этого можно сделать вывод, что критерии отнесения слов к условным заимствованиям не были достаточно четкими.

6. Работа над словами с искажениями

Обучающее множество, на котором мы тренировали классификаторы отмечать слова с искажениями, изначально состояло из 700 слов. Такой размер первичной выборки обусловлен тем, что она составлялась вручную и состояла преимущественно из слов, случайным образом найденных нами в данных. Для того, чтобы составить обучающее множество и определить, какие типы искажений есть в данных, мы случайным образом извлекали из датасета выборки и искали среди них примеры слов с искажениями. Также мы добавили в обучающее множество относящиеся к “олбанскому языку” слова из материала “Берлога веблога. Введение в эрратическую семантику”[20], поскольку они также отражают некоторые стратегии изменения написания слова.

В рамках этой работы мы сосредоточились на фонологических искажениях, под которыми мы подразумеваем замену одних букв другими с нарушением орфографических и/или фонологических норм. Этим целевые для нас искажения отличаются от механических, то есть таких, которые отличаются от словарных форм только изменением количества букв в словах (например, умножение букв – “маааамооочкаааа”, вставка букв – “воут”, перестановка букв – “роисся”). Важно отметить, что нас в первую очередь интересуют именно эрративы, а не случайно возникшие ошибки или опечатки. Однако проблема поиска отличий между ошибками и намеренными искажениями не рассматривалась в рамках этой работы и может рассматриваться как продолжение исследования в этой области.

Во время изучения датасета мы обнаружили фонологические искажения, которые можно объединить в следующие группы.

Таб. 2. Примеры эрративов

Виды	**Примеры подвидов**	**Примеры слов**
замена одних согласных другими	“ц”/”цц” вместо “ться”/”тся”/”ся”/”ть”/”чь” на конце глаголов	*принюхиваца*, *состариццо*, *розвиватца*
	“л”/”в”/”й”/”д” вместо “р”	*секлетали, халясё*, *выздовавливай, вывирай, аймагедон, вздослая*

[20] Берлога веблога. Введение в эрратическую семантику // Архивы форума "Говорим по-русски" URL: http://www.speakrus.ru/gg/microprosa_erratica-1.htm (дата обращения: 04.04.2020).

озвончение глухих согласных и оглушение звонких согласных		*держизь, чазовой, бодмышки, апсалютно, афтомат, вофчика*
переход от смычных к фрикативным звукам и наоборот	“ш”/”щ” вместо “ч”	*романтишный, палощкою, канеш*
	“ч”/”сч” вместо “щ”	*вапче, вобсче*
смещение от нёбных звуков к зубным	“з” вместо “ж”	*бозе, лазденя*
замена одних гласных другими	“у”/”ю” вместо “е”/”и”	*поплюваться, сурьезный, чумадан, денюшек*
	“ы” вместо “и”	*спецыальность, ашыпка, гваздыка, дыстрыбутар*
	“а” вместо “о”	*пампездна, одназначна, кисленька, перанаснога, атрежу*
	“э” вместо “е” и “е” вместо “э”	*тэрпеть, цэлых, крэмль, еротику, експеримент*
	“о” вместо “е”	*суперзачот, зачотненько, зажогши*
замена йотированных гласных на «й» и обычный гласный	“йо”, “йу” и “йа” вместо “ё”, “ю” и “я”	*йожиг, йуный, йазва*
сокращения, похожие на измененную транскрипцию слова		*тарищи, спрашали, ващет*

По приведенным выше примерам можно заметить, что некоторые слова из обучающей выборки и, соответственно, из наших данных сочетают в себе несколько видов искажений сразу, что отображает стратегии написания слов, используемых в интернете: зачастую пользователи искажают не одну букву в слове, а сразу несколько.

В качестве обучающего множества словарных, то есть неискаженных, слов мы использовали словарь Зализняка со словоформами. Случайным образом из словаря была составлена первичная выборка в размере 1000 слов. Эти слова мы отметили как нецелевой класс для поиска и объединили со словами целевого класса (выборкой

искажений). Датасет сильно несбалансированный, но оценить точное соотношение искажений и неискажений в нем вручную или до запуска первичной модели не представлялось возможным, поэтому мы остановились на таких числах.

Разбив выборку на обучающую и тестовую, мы запустили логистическую регрессию в первый раз и получили на тесте показатель f-меры, равный 0.70. Такое значение свидетельствовало, что в целом модель выбирала слова с искажениями неслучайным образом. Просмотр слов, отобранных с помощью классификатора, показал, что среди них присутствуют не только определенные нами виды фонологических искажений, но и следующие слова.

Таб. 3. Примеры ложных выявленных эрративов.

Виды	**Примеры слов**
слова с добавленными буквами на конце	*грозитъ, режиссеромъ*
склееные слова	*теплыевещички, центредеревни, искреннесоветую*
слова с повторяющимися слогами и буквами	*аххххахахахахахаха*
слова с креативной морфологией	*психургия, масквaляндия, натарфаретил, пьятнеца, ляпатуська, одноинститутница*, *деблохизация*, *девочкогусь*
слова с пропущенными буквами	*экспертнго*, *невозжно*
слова, в которые попала соседняя на клавиатуре компьютера или телефона буква	*нвкопилось*
слова с неграмматичной эпентезой	*дорогра*
слова с метатезой	*кбалуках*

Под “словами с креативной морфологией” мы подразумеваем окказионализмы и слова, образованные с помощью словарных морфем, скомбинированных нестандартным образом. Эти слова не являлись целевым классом для нашего исследования, однако благодаря работе модели мы увидели примеры таких слов. Изучение этого класса может расширить начатую работу в дальнейшем.

Среди наших данных встречались слова, которые явно были написаны со случайными ошибками и опечатками, некоторые из них модель разметила как намеренные искажения. Получив первые результаты, мы также выявили в данных

слова, которые не являются словарными, но и не могут быть отнесены ни к одному из заданных нами классов (примеры таких слов: *“саклаунын”*, *“акасакала”*, *“кёнке”*, *“ытгытыя”*). Часть таких слов может принадлежать другим языкам.

Благодаря первым полученным результатам мы увеличили обучающую выборку искажений. На момент повторного запуска модели число эрративов составило 1035. Мы также использовали случайную выборку слов из словаря Зализняка и добавили к ней 2205 несловарных словоформ, которые были ошибочно выявлены как искажения после первого запуска регрессии. Затем мы повторно запустили логистическую регрессию. В результате мы подняли значение f-меры на тестовой выборке до 0.73.

Мы решили попробовать использовать другие алгоритмы в качестве классификаторов, а именно Random Forest и метод стохастического градиентного спуска, и посмотреть, поднимется ли значение f-меры до 0.80 при использовании этих методов. При запуске метода Random Forest мы получили значение f-меры 0.76, а при использовании стохастического градиента f-мера составила 0.79. Таким образом, из трех использованных нами алгоритмов согласно f-мере лучше других сработал метод стохастического градиента. По этой причине для выявления всех искажений на всем датасете мы использовали именно его.

Всего с помощью метода градиентного спуска мы выявили 2 703 611 слов с различными видами искажений.

В новой разметке данных среди ошибочно выявленных как искажения слов мы обнаружили следующие.

Таб.4. Примеры ложных выявленных эрративов.

Виды	**Примеры слов**
слова с экспрессивными аффиксами	*алиночка, верзилища, виндуха, гагашенька*
слова с креативной морфологией	*вискари, витаминьтесь, вображульство, воркаголики, вчерасть, гламурьё, гуглирует*
слова, состоящие из склеенных словарных слов	*айхочухочухочухочу, вамчемнибудьпомочь, вирусимунодефицитачеловека*
слова со слогами, которые поменяли местами	*ведьмежонок*
звукоподражательные слова и слова с размноженными слогами	*аэуа, ахаха, гыгыгыгы, дамдаридаридаридамдам*

имена собственные	абдурахманов, вегабизнесконсалтинг, виктордмитрич, *дашунь*
аббревиатуры	*вгтрк, гибдэдэ*

На данный момент эти слова находятся в нашей базе искажений. В дальнейшей работе можно очистить базу искажений от слов, примеры которым даны в этой таблице. Мы приняли решение оставить эти слова в базе, поскольку в части из них можно проследить стратегии искажения слов, не изученные и не найденные нами. Например, механические искажения, которые не являлись целевым классом для данной работы, или же несловарные сочетания словарных слов с аффиксами могут представлять ценность для других исследователей, которые смогут воспользоваться полученной в результате нашей работы базой.

По результатам работы на случайной выборке, состоявшей из 200 слов, был получен коэффициент альфа Криппендорфа. Для слов, относящихся к фонологическим искажениям, он составил 0.3919. Аннотаторы также оценили принадлежность слов к классу “креативной морфологии”, значение коэффициента в этом случае составило 0.4278. Такие показатели свидетельствуют, что в целом аннотаторы разошлись во мнениях относительно принадлежности слова к классу искажений. Это может объясняться как вариативностью в человеческом понимании того, какие слова относятся к искажениям, так и недостаточно детальными критериями. В дальнейшем предполагается продолжить эксперименты по выявлению эрративов с определением и использованием более конкретных критериев принадлежности к искажениям.

7. Работа над экспрессивными формами

Для множества слов с экспрессивной окраской, на котором классификатор обучался находить аналогичные экспрессивно окрашенные слова, нам удалось собрать 441 слово, содержавшее один или несколько экспрессивных аффиксов, отобранных нами для этого исследования. Тренировочное множество частично состояло из экспрессивных форм, найденных нами в случайных выборках из исходных данных (под экспрессивными формами мы понимаем слова, содержащие аффиксы, придающие экспрессивную окраску слову). Часть обучающей выборки составили слова, которые были отобраны из ряда источников, в той или иной степени посвящённых аффиксам, придающим словам экспрессивную окраску. Это материалы из учебника Розенталя, Голуб и Теленковой[21], с сайта “Словород”[22] и исследования Покровского[23], текст

[21] Розенталь Д.Э., Голуб И.Б., Теленкова М.А. Современный русский язык. М.: Айрис-Пресс, 2002.
[22] Словарь русских суффиксов на Словороде. // Словород. Сайт о происхождении и образовании русских слов URL: http://www.slovorod.ru/russian-suffixes.html (дата обращения: 17.03.2020).
[23] Покровский В. Оценочность суффиксов в русской речи // Работа на XIV научно-практическую конференцию школьников «Старт в науку». Прокопьевск: 2011.

“Значение суффиксов для русской литературы”[24], работы Саримовой[25], Медведевой[26], Пучковой[27], Фуфаевой[28]. Также часть примеров мы выбрали из словаря Зализняка. Ниже приведены примеры слов, основываясь на которых, модель должна была искать слова нужной нам целевой группы:

Таб. 5. Примеры экспрессивных форм

Виды форм	**Примеры аффиксов**	**Примеры слов**
имена существительные со 139 экспрессивными суффиксами	“*ик*"	*хлюпик*, *конвертик, Олик*
	"*к*"	*дочурка*, *ладошки*, *гринписьки*
	"*ас*"	*мемасы*, *пивас*, *Юрасик*
	"*ичк*"	*сестричка*, *косичка*, *водичка*
	"*яг*"	*умняга*, *миляга*, *крутяга*
наречия и имена прилагательные с 10 суффиксами	"*еньк*"	*скрытненько*, *осторожненько, утробыстренько*
	"*оньк*"	*легонько*, *легонький*, *мягонький*
	"*ющ*"	*вкуснющим*
	"*ехоньк*"	*близехонько*, *белехоньким*, *смирнехонький*
	"*онечк*"	*тихонечко*
глаголы с суффиксом "*к*"	"*к*"	*тетёшкать*, *пиздявкать*
междометия с суффиксами	"*ечк*"	*божечки*

[24] Значение суффиксов для русской литературы - сочинение // Антисочинение URL: http://antisochinenie.ru/сочинения/Другие_сочинения_по_зарубежной_литературе/Значение_суффиксов_для_русской_литературы (дата обращения: 15.03.2020).
[25] Русский язык и культура речи : лекции / сост.: Р.Р. Саримова. – Нижнекамск : Нижнекамский химико-технологический институт (филиал) КГТУ, 2009.
[26] Медведева К. М. Семантика эмоционально-экспрессивных суффиксов квалитативных форм русских антропонимов // Молодой ученый. 2013. №7 (54).
[27] Пучкова С.А. Семантико-грамматическая классификация оценочных слов в рассказах М. Зощенко // Язык. Культура. Коммуникации. 2018. №2.
[28] Фуфаева И.В. Экспрессивные диминутивы в условиях конкуренции с нейтральными существительными: на материале русского языка: дис. ... канд. фил. наук: 10.02.01. М., 2018.

	"*ошеньк*"	*фигошеньки*
слова, использующие экспрессивные префиксы	"*супер*"	*супермилым*, *супермужчину*, *супермышц*
	"*мега*"	*мегадоклад*, *мегатоталитарная*
	архи	*архичасто*

Среди примеров, встречающихся в обучающей выборке, есть слова с одним (“*кисуля*”, “*сламечка*”, “*пачпортинка*”) и двумя (“*справулечек*”, “*хняшечка*”, “*совятки*”) аффиксами, создающими экспрессивную окраску.

В рамках данного исследования мы не разделяли слова с префиксами, содержащими экспрессивную окраску (“*супермилым”*) и без нее (“супервайзер”, “мегафон”). Однако слова с такими аффиксами, но без экспрессивной окраски, не были добавлены в тренировочное множество. В дальнейшем предполагается расширить работу над такой лексикой за счет исследования данного разделения.

Для обучающего множества слов без экспрессивной окраски мы использовали словоформы из словаря Зализняка. Так как количество слов с экспрессивной окраской в исходном датасете должно быть меньше, чем количество неэкспрессивных форм, мы решили отобрать большее количество слов без экспрессивной окраски для обучающего множества: из всего списка была взята случайная выборка, из которой были вручную отобраны 2955 слов, не имеющих экспрессивных аффиксов. Эти слова были отмечены нами как нецелевые для поиска, после чего мы объединили их со словами с экспрессивной окраской. На объединенном множестве слов мы обучили классификаторы – логистическую регрессию и стохастический градиентный спуск – находить слова с экспрессивной окраской, чтобы отделять такие слова от всех остальных.

Для каждого слова было получено векторное представление с помощью предобученной модели fastText. Далее эти представления были использованы для классификатора, использующего логистическую регрессию. Если классификатор присваивал слову класс 1, то оно сохранялось в отдельный файл.

Запустив логистическую регрессию на этом обучающем множестве, мы получили f-меру, равную 0.76, на тестовой выборке. Этот показатель демонстрирует, что модель в основном выбирала экспрессивно окрашенные слова неслучайно.

Однако с помощью классификатора, использующего стохастический градиентный спуск, удалось получить f-меру, равную 0.9, на тестовой выборке, что является существенно лучшим результатом. Таким образом, для работы с полным датасетом нами использовался именно этот классификатор. Просмотр извлечённых с помощью стохастического спуска слов с экспрессивной окраской показал, что, помимо слов, относящихся к целевому классу исследования, были и ошибочно отобранные слова.

Таб.6. Примеры ложных экспрессивных форм.

Виды	**Примеры слов**
слова из 2-3 букв	*ХС*, *ТЩЕ*, *Чиё*
"склееные" слова	*умненькийстрашненькийбезтелаилица*, *простонепередатьсловаааааамиииииии*, *огурцовпомидоровкапустымандаринов*
слова, в которых повторяются слоги и буквы	*мвахахахахахаахахахахахахаххахаха*, *вооооооооооооооооооооооооооот*, *домоооооооооооооооооооойййййййййй*
несловарные англицизмы	*лайфтайм*, *бэтмансити*
слова с опечатками	*магазаны*
слова с намеренными/ненамеренными искажениями	*чирикале*
имена собственные	*варегемом*, *шиховски*

Кроме того, мы получили ряд несловарных слов, смысл которых неясен, но и к экспрессивным формам их отнести нельзя (примеры: “*лунды*”, “*мопара*”, “*сорце*”, “*напотагама*”). Часть таких слов может относиться не к русскому языку.

Также среди выбранных классификатором форм есть достаточно большое количество иных слов, не относящихся ни к экспрессивным формам, ни к одному из целевых классов. Такие слова не имеют ярко выраженных характерных особенностей, поэтому их достаточно сложно классифицировать.

Нетрудно заметить, что модель ошибочно отмечает как экспрессивные слова, которые относятся к другим целевым категориям. Для получения более точной базы экспрессивных слов в дальнейшей работе можно попробовать многоклассовые модели, которые могут отделить подобные слова от экспрессивных или определить формы как относящиеся к нескольким классам одновременно.

Исходя из полученных результатов было решено игнорировать слова с количеством символом менее 4 и более 18. Количество извлекаемых слов в результате увеличилось с 805241 до 1071964.

Среди тех примеров, которые мы извлекали с помощью случайной выборки, количество слов, не являющихся экспрессивно окрашенными, достаточно, чтобы сделать вывод, что модель необходимо улучшать (например, с помощью увеличения обучающего множества).

В дальнейшем можно расширить нашу работу исследованием обсценной лексики, которая встречается в интернете.

8. Заключение

Итогом работы на текущий момент можно считать следующее. Применение предобученных моделей эмбеддингов, использующих при обучении информацию о внутренней структуре слова (подсловах), оказалось очень эффективным, о чем свидетельствуют показатели f-меры для каждого целевого класса. Исходя из полученных результатов классификации, можно говорить, что сжатые признаки, содержащиеся в векторных представлениях отдельных слов, несут некоторую долю морфологической информации, о чем свидетельствует успешная работа классификаторов на словах, в которых основными различительными признаками являлись именно морфологические характеристики.

Также можно говорить о существовании линейной зависимости между данными признаками, о чем свидетельствует эффективность линейных моделей классификации.

Другим итогом стало то, что, как выяснилось, качество самой модели при обучении не обязательно переходит в сравнимое качество при разметке аннотаторами. Это верно даже для того случая, в котором, по нашим предположениям, должно быть минимум разногласий – для заимствованных слов. Одной из причин, по которым возникли такие разногласия, может быть то, что при изначальном разделении на классы критерии принадлежности слов к целевым классам были более вариативны, чем это необходимо. Из-за этого целевые классы могли получиться слишком широкими и открытыми для интерпретации. В дальнейшем можно попробовать определить более жесткие критерии для отбора слов. Другой причиной может быть изначально сильная несбалансированность датасета, которую было трудно оценить ввиду отсутствия разметки по классам в первоначальных данных. Ввиду того, что собрать вручную оптимальную выборку для обучения достаточно затруднительно, распределение классов в обучающей и тестовой выборке могло сильно отличаться от распределения классов в основных данных.

Мы попробовали использование бинарных классификаторов, так как опирались на предположение, что одно слово может принадлежать только одному классу. По всей видимости, верно то, что каждое слово может относиться больше, чем к одному из обозначенных нами классов. Поэтому для улучшения качества модели, повышения согласия между аннотаторами и создания более точной базы интернет-лексики целесообразно использовать мультилейбловую (multilabel) модель, то есть модель, которая выдает для каждого класса независимую вероятность и которая способна предсказывать вероятность отнесения слова к нескольким классам одновременно. Предполагается, что аннотаторы могут больше соглашаться с тем, что слово относится одновременно к нескольким классам, а база слов будет точнее отображать стратегии использования и изменения слов из интернета.

9. Список литературы и источников

1. Antidictionary: database of out-of-dictionary words from the Russian Internet // AINL 2018 – Agenda. Poster and demo session. Posters URL: https://ainlconf.ru/2018/agenda#program-demo-and-posters (дата обращения: 26.06.2020).
2. Вахранев А.Ю. Намеренное нарушение языковой нормы: исследование на основе языковых корпусов // Пространство научных интересов:иностранные языки и межкультурная коммуникация – современные векторы развития и перспективы: сборник статей по результатам IV научной межвузовской конференции молодых ученых 11.04.2019 г. (ДИЯ НИУ ВШЭ). Отв. редактор Е.Г. Кошкина. М.: ООО «Буки Веди», 2019. С. 32-40.
3. Зализняк А. А. Механизмы экспрессивности в языке // Ю.Д. Апресян и др. (ред.). Смыслы, тексты и другие захватывающие сюжеты: Сб. ст. в честь 80-летия Игоря Александровича Мельчука. — М.: ЯСК, 2012. — С. 650–664.
4. Карасева, А.И. Роль и функции эрратива в интернет-сленге // Мониторинг общественного мнения: экономические и социальные перемены. 2008. №2(86). С. 129-140.
5. Косенко Е.И. Анализ функционирования экономических заимствований в современных российских СМИ // Вестн. Сев. (Арктич.) федер. ун-та. Сер.: Гуманит. и соц. науки. 2017. № 2. С. 107-113. DOI: 10.17238/issn2227-6564.2017.2.107
6. Кудинова, Т.А. Язык Интернет и жаргон падонков // Актуальные проблемы филологии и педагогической лингвистики. 2010. С. 336-340.
7. Медведева К. М. Семантика эмоционально-экспрессивных суффиксов квалитативных форм русских антропонимов // Молодой ученый. 2013. №7 (54).
8. Орехов Б.В. Суперминимум и нанодержава: префиксоиды в языке интернета // Современный русский язык в интернете / ред. Я. Э. Ахапкина, Е. В. Рахилина. — М.: Языки славянской культуры, 2014. С. 281-290.
9. Покровский В. Оценочность суффиксов в русской речи // Работа на XIV научно-практическую конференцию школьников «Старт в науку». Прокопьевск: 2011.
10. Пучкова С.А. Семантико-грамматическая классификация оценочных слов в рассказах М. Зощенко // Язык. Культура. Коммуникации. 2018. №2.
11. Розенталь Д.Э., Голуб И.Б., Теленкова М.А. Современный русский язык. М.: Айрис-Пресс, 2002.
12. Русский язык и культура речи : лекции / сост.: Р.Р. Саримова. – Нижнекамск : Нижнекамский химико-технологический институт (филиал) КГТУ, 2009.

13. Словарь языка Интернета.RU / Кронгауз М. А., Пиперски А. Ч., Сомин А. А., Черненко Ю. А., Мерзлякова В. Н., Литвин Е. А., Под ред. Кронгауза М. А. М.: АСТ-Пресс Книга, 2016.
14. Современный русский язык в интернете / Под ред. Я. Э. Ахапкина, Е. В. Рахилина. М.: Языки славянской культуры, 2014.
15. Феногенова А.С., Карпов И.А., Казорин В.И., Лебедев И.В. Сравнительный анализ распределения англицизмов в русских текстах социальных медиа // Сборник Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 1 — М.: Изд-во РГГУ, 2017. С. 65-74.
16. Фуфаева И.В. Экспрессивные диминутивы в условиях конкуренции с нейтральными существительными: на материале русского языка: дис. ... канд. фил. наук: 10.02.01. М., 2018.

Интернет-ресурсы

1. Importance of character n-grams // Word representations. FastText Docs URL: https://fasttext.cc/docs/en/unsupervised-tutorial.html#importance-of-character-n-grams (дата обращения: 15.09.2019).
2. MyStem // Яндекс URL: https://yandex.ru/dev/mystem/ (дата обращения: 14.04.2020).
3. Берлога веблога. Введение в эрратическую семантику // Архивы форума "Говорим по-русски" URL: http://www.speakrus.ru/gg/microprosa_erratica-1.htm (дата обращения: 04.04.2020).
4. Другие корпуса // Национальный корпус русского языка URL: http://www.ruscorpora.ru/new/corpora-other.html (дата обращения: 14.04.2020).
5. Задача: извлечь ключевые выражения из текста на русском языке. NLP на Python // Habr URL: https://habr.com/ru/post/468141/ (дата обращения: 20.03.2020).
6. Значение суффиксов для русской литературы - сочинение // Антисочинение URL: http://antisochinenie.ru/сочинения/Другие_сочинения_по_зарубежной_литературе/Значение_суффиксов_для_русской_литературы (дата обращения: 15.03.2020).
7. О проекте // Генеральный Интернет-Корпус Русского Языка URL: http://www.webcorpora.ru (дата обращения: 14.04.2020).
8. Словарь // Дьяков А.И. Словарь англицизмов русского языка URL: http://anglicismdictionary.ru/Slovar (дата обращения: 10.09.2019).
9. Словарь Зализняка oDict.ru // Открытый грамматический словарь русского языка URL: http://odict.ru (дата обращения: 10.09.2019).
10. Словарь русских суффиксов на Словороде. // Словород. Сайт о происхождении и образовании русских слов URL: http://www.slovorod.ru/russian-suffixes.html (дата обращения: 17.03.2020).
11. Статические модели // RusVectōrēs: семантические модели для русского языка URL: https://rusvectores.org/ru/models/ (дата обращения: 04.04.2020).