КЛАССИФИКАЦИЯ ОТЗЫВОВ IMDB ПО ЭМОЦИОНАЛЬНОМУ ОКРАСУ ТЕКСТА ПРИ ПОМОЩИ МЕТОДОВ БОЛЬШИХ ДАННЫХ

Щербинина Юлия Олеговна

студент, Московский политехнический университет РФ, г. Москва

Суворов Станислав Вадимович

научный руководитель, канд .экон. наук, профессор кафедры «Прикладная информатика» Московский политехнический университет, РФ, г. Москва

CLASSIFICATION OF IMDB REVIEWS BY THE EMOTIONAL COLOR OF THE TEXT USING BIG DATA METHODS

Julia Shcherbinina

Student of Moscow Polytechnic University, Russia, Moscow

Suvorov S.V.

Scientific adviser, Ph.D. Professor of the Department of Applied Informatics Moscow Polytechnic University, Russia, Moscow

Аннотация. Классификация текста — важная проблема обработки естественного языка (Natural Language Processing, NLP). Одна из основных трудностей при анализе — это многозначность, способная возникнуть на всех этапах работы. Обычно такие данные удаляются при помощи контекста, учитывая особенности в построении конструкций словосочетаний и предложений в языке. Другая проблема связана с тем, что методы анализа текстовой информации напрямую зависят от языка, жанра, предметной области. Анализ художественного текста не то же самое, что анализ новости или текста из социальных сетей, поэтому всегда требуется дополнительная настройка для исследуемых данных.

Abstract. Text classification is an important problem in Natural Language Processing (NLP). One of the main difficulties in analysis is the ambiguity that can arise at all stages of work. Typically, such data is removed using the context, taking into account the peculiarities in the construction of word combinations and sentences in the language. Another problem is related to the fact that the methods of analyzing textual information directly depend on the language, genre, and subject area. Analysis of fictional text is not the same as analysis of news or text from social networks, so additional adjustment is always required for the data being examined.

Ключевые слова: Нейролингвистическое программирование, классификация текста, отзывы, нейронные сети, Большие данные.

Keywords: Neuro-Linguistic Programming, text classification, feedback, neural networks, Big data.

Актуальность исследования заключается в том, что большая часть информации в Интернете не имеет структуризации, а значит всё сильнее возрастает потребность в задачах её классификации.

Объектом исследования являются отзывы пользователей и критиков на продукты кинематографа с сервиса IMDB.

Предметом исследования является метод обработки больших данных, направленный на выявление списка слов, характеризующий принадлежность отзывов к определенным группам.

Цель состоит в том, чтобы провести анализ данных киноиндустрии методами проектной аппроксимации для того, чтобы в дальнейшем автоматически классифицировать текстовые документы по одной или нескольким предопределенным категориям. Сложность обучения модели классификации текста может быт связана с проблемами исходных данных

В вопросах классификации документов, таких как тональность классификации, выбор представления документа обычно важнее, чем выбор классификатора. Задачи текстового представления направлены на отображение текстов переменной длины в векторы фиксированной длины, чтобы быть допустимыми входными данными для классификатора.

Для достижения вышеуказанной цели, необходимо решить следующие задачи:

- Изучить информацию об отзывах в киноиндустрии.
- Составить векторное пространство для обучения нейронной сети.
- Сделать анализ данных киноиндустрии.
- Проанализировать полученные результаты.

Ранние подходы машинного обучения для классификации текстов были основаны на извлечении признаков мешка слов с последующим использованием контролируемого классификатора. Недавно были введены модели рекуррентных и конволюционных нейронных сетей для использования порядка слов и грамматической структуры.

Недавние улучшения в моделях NLP во многом основывались на парадигме предварительного обучения и тонкой настройки. При этом языковая модель сначала предварительно обучается на массивных текстовых корпусах, а затем выполняется точная настройка по последующим задачам. Производительность модели варьируется в зависимости от поставленных задач и количества доступных обучающие примеры. Однако, на практике существует слишком большое количество доменов, задач и языков, и масштабирование до нового проблемного пространства потребует дополнительных помеченных данных. Это ведет к важной области исследования — обучению по нескольку раз, которое предполагает доступ только к небольшому количеству помеченных примеров.

Далее будет проведено исследование с использованием исходных данных и применены методы обработки естественного языка, чтобы определить отношение писателя к конкретному продукту киноиндустрии. Вообще говоря, анализ тональности — это форма классификации текстовых документов по многочисленным группам. В большинстве случаев нужно только классифицировать документы на положительные и отрицательные классы. Кроме того, существуют различные методы анализа настроений, которые могут помочь измерить настроения. Эти методы включают методы лексических подходов и методы контролируемого машинного обучения. Модели машинного обучения более популярны, потому что лексические подходы, основанные на семантике слов, используют заранее определенный список положительных и отрицательных слов для извлечения тональности новых документов. Создание этих предопределенных списков отнимает много времени, и мы не можем создать уникальный лексический словарь, который можно было бы использовать в каждом отдельном контексте. С ростом популярности социальных сетей постоянно создаются огромные наборы больших данных обзоров, блогов и каналов социальных сетей, посвященных оценке тех или иных вещей, в том числе и кино. Методы больших данных используются в доменах приложений, которые собирают и хранят огромные объемы данных. Растущий объем

данных, интенсивные технологии и увеличение ресурсов хранения данных развивают науку о больших данных. Основная концепция аналитики больших данных - извлечение значимого шаблона из огромного количества данных. Для больших данных нужны специальные методы, которые можно использовать для извлечения закономерностей из огромного количества данных. Глубокое обучение имеет эту возможность предоставить решение для решения проблемы обучения и анализа данных, которая существует в огромном количестве данных, а также они лучше обучаются сложным шаблонам данных. Существуют и другие проблемы с большими данными, такие как внедрение домена и потоковая передача данных, с которыми приходится бороться крупномасштабным моделям глубокого обучения для аналитики больших данных. Концепции и методы анализа настроений, которые могут помочь извлекать информацию из этих областей, становятся все более важными, поскольку предприятия, организации и отдельные лица стремятся лучше использовать свои большие данные для определения настроений целевой аудитории.

Методы обучения с учителем подвержены проблеме переобучения. Переобучение — это явление, когда построенная модель хорошо классифицирует примеры из обучающего набора, но классифицирует любые другие примеры относительно плохо. Это происходит потому, что в процессе обучения модель выявляет некоторые закономерности в обучающей выборке, которые отсутствуют в общей популяции. Как бороться с переобучением зависит от конкретных методов и моделей.

Машинное обучение в анализе тональности сводится к обычной классификации.

Таким образом, чтобы правильно классифицировать текст необходимо совершить следующие шаги:

1. При помощи словаря токенизировать текст;
2. Преобразовать полученный словарь в векторы значений;
3. Математическими методами вычислить к какому классу принадлежит каждый из векторов.

Для более успешной классификации, датасет разделен на три области, позитивную, негативную и нейтральную. Больше половины занимает негативные отзывы (Рисунок 1)

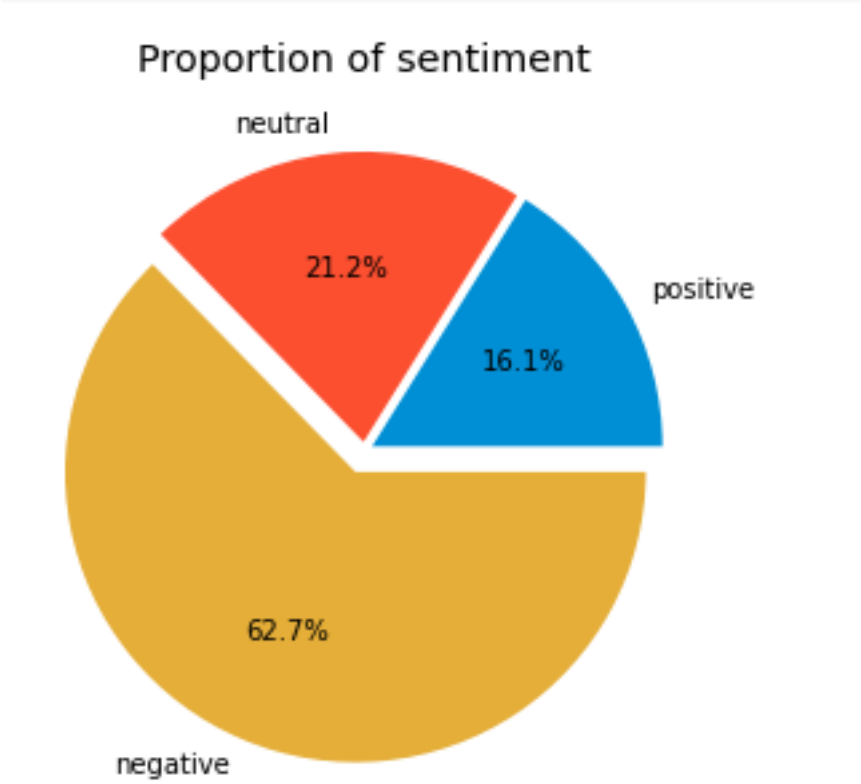

Рисунок 1. Процентное соотношение эмоционально окрашенного текста

Благодаря этому у рекуррентной нейронной сети будет достаточно данных, как для обучения, так и тестирования результатов обучения для дальнейшей корректировки.

Используемый в исследовании набор данных с 360000 образцами текста, разбитыми на две колонки одна самим текстом, а другая с его оценкой: позитивной или негативной.

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

Рисунок 2. Набор исходных данных

Понять в каком направлении оценить текст (как позитивный, как негативный или как

нейтральный) можно по его оценкам на самом сервисе «IMDB». Если оценка больше шести баллов, то отзыв считается положительным, если меньше четырех, то отрицательным, в остальных случаях (от четырех до шести) отзыв будет нейтральным.

Сами отзывы содержат множество местоимений, союзов и частиц, они же и являются самыми часто встречающимися словами, для успешного анализа данных необходимо их исключить, чтобы не зашумлять выборку данных. Для этого составляются три функции очищения текста, куда и добавляются все ненужные части речи. В первой убираются все html-метки из текста, такие как перенос на новую строку, жирный шрифт или курсив, второй функцией приводится весь текст к нижнему регистру и убирает все абзацы или отступы из текста, третья функция отвечает за удаление всех точек и запятых из отзывов для лучшего обучения модели выбранным методом.

Перед тем, как начать полноценно обрабатывать датасет и обучать модель с помощью выбранных методов необходимо провести предварительную обработку всех данных. А именно токенизировать их, а также нормализовать.

Преобразование набора данных в числовые тензоры — обычно называется векторизацией. Поскольку модели глубокого обучения не принимают в качестве входных данных необработанный текст: они работают только с числовыми тензорами, поэтому векторизация проводиться в обязательном порядке. Есть несколько способов сделать это, самый распространённый это при помощи библиотеки NLTK и метода TD-IDF.

Токенизация — означает разбиение текста на составляющие, в данном случает разбивать отзыв на несколько связанных друг с другом слов. Токены включают в себя не только слова, но также и пунктуацию. Поэтому после токенизации необходимо сделать чистку: убрать знаки препинания и незначащие слова (например, предлоги). Такие действия имеют свой термин под названием «Нормализация».

Нормализация — означает, что нам необходимо привести все буквы в тексте к единому регистру, убрать союзы, а также удалить все слова, которые будут создавать лишний шум.

Нейронные сети обладают большей эффективностью, чем метод Наивного байесовского классификатора, однако в других метриках (легкость и простота) уступают.

Для правильной оценки работы полученной модели разделим модель на тестовую часть и проверочную. Нужно разделять обучающие и тестовые наборы данных перед любой подготовкой данных. Это означает, что какие-либо знания в данных в наборе тестов, которые могут помочь лучше подготовить данные (например, используемые слова), недоступны при подготовке данных, используемых для обучения модели.

Поскольку компьютерная модель не воспринимает естественный язык, то необходимо представить все слова в отзыве в числовом виде. Для этого векторизуется каждый отзыв, который представляется его в виде вокубуляра состоящего из уникальных лексем — даётся определение выше.

Первое, что необходимо сделать — это создать общий список, в котором будут находиться отзывы и связанные с ним классы, вместе с их метками, после чего создается единый вокубуляр.

После чего нужно соединить их, используя 15 000 самых часто используемых слов. В итоге получается вокубуляр, который будет состоять из 20 000 самых часто используемых слов.

Преобразование набора данных в числовые тензоры — обычно называется векторизацией. Поскольку модели глубокого обучения не принимают в качестве входных данных необработанный текст: они работают только с числовыми тензорами, поэтому векторизация проводиться в обязательном порядке. Есть несколько способов сделать это, самый распространённый это при помощи библиотеки NLTK и метода TD-IDF.

Токены включают в себя не только слова, но также и пунктуацию. Поэтому после токенизации

необходимо сделать чистку: убрать знаки препинания и незначащие слова (например, предлоги). Такие действия имеют свой термин под названием «Нормализация».

В результате получается отфильтрованный список слов, который подходит для дальнейшей обработки.

После токенизации текста необходимо построить модель нейронной сети, которая и будет заниматься классификацией. Для этого необходимо создать несколько слоев многословной нейронной сети, при том, что во водные слои необходимо подавать тексты определенного размера, который подбирается экспериментальным путем. Вышеупомянутый слой принимает двумерные целочисленные тензоры формы и по крайней мере два аргумента: количество возможных токенов и размерность вложений. Наконец, он возвращает трехмерный тензор формы с плавающей запятой, который теперь может обрабатываться нейронной сетью. Теперь можно переходить к следующему этапу:

Наивный Байес высчитывает вероятность принадлежности каждого отзыва к одному из двух классов: положительного или отрицательного. Для этого, перемножая условные вероятности появления всех известных слов в отзыве, при условии их принадлежности к тому или иному классу.

Методом подбора находится оптимальный гиперпараметр α (alpha), который служит для того, чтобы сглаживать выбросы и шум в модели. Также, если определенное слове не встречалось в выборке из 10 000 обучающих слов, то его условная вероятность будет равно нулю, следовательно обнуляет вероятность быть определенной к какому-либо классу. Чтобы избежать большого количества ошибок, необходим этот гиперпараметр α, который по умолчанию добавляет некоторое значение к условной вероятности

Как правило, информация о настроениях передается с помощью прилагательных или, более конкретно, определенными сочетаниями прилагательных с другими частями речи. Эту информацию можно получить, добавив такие функции, как последовательные пары слова – биграммы, или даже тройки слов – триграммы. Далее определяется количество слоев в нейронной сети: при реализации модели Keras Sequential все дело в наложении слоев. Слои нейронной сети могут принимать несколько аргументов, но определив 15, которые представляют собой количество скрытых единиц в слое (должно быть положительным целым числом и представлять размерность выходного пространства) и процент отсева слоя. Отсев — один из наиболее эффективных и наиболее часто используемых методов регуляризации для нейронной сети, который заключается в случайном отключении скрытых модулей во время обучения, таким образом сеть не полагается на 100% на все свои нейроны, а вместо этого заставляет себя находить более значимые шаблоны в данные, чтобы увеличить показатель, который вы пытаетесь оптимизировать.

Алгоритмом оптимизации был выбран Adam, поскольку фактический размер шага, сделанный Адамом на каждой итерации, приблизительно ограничен гиперпараметром размера шага. Это свойство добавляет интуитивное понимание к предыдущему неинтуитивному гиперпараметру скорости обучения.

Размер шага правила обновления Адама инвариантен к величине градиента, что очень помогает при прохождении через области с крошечными градиентами (например, седловые точки или овраги).

Adam был разработан, чтобы объединить преимущества Adagrad, который хорошо работает с разреженными градиентами, и RMSprop, который хорошо работает в онлайн-настройках. Обладая и тем, и другим, можно использовать Адама для более широкого круга задач.

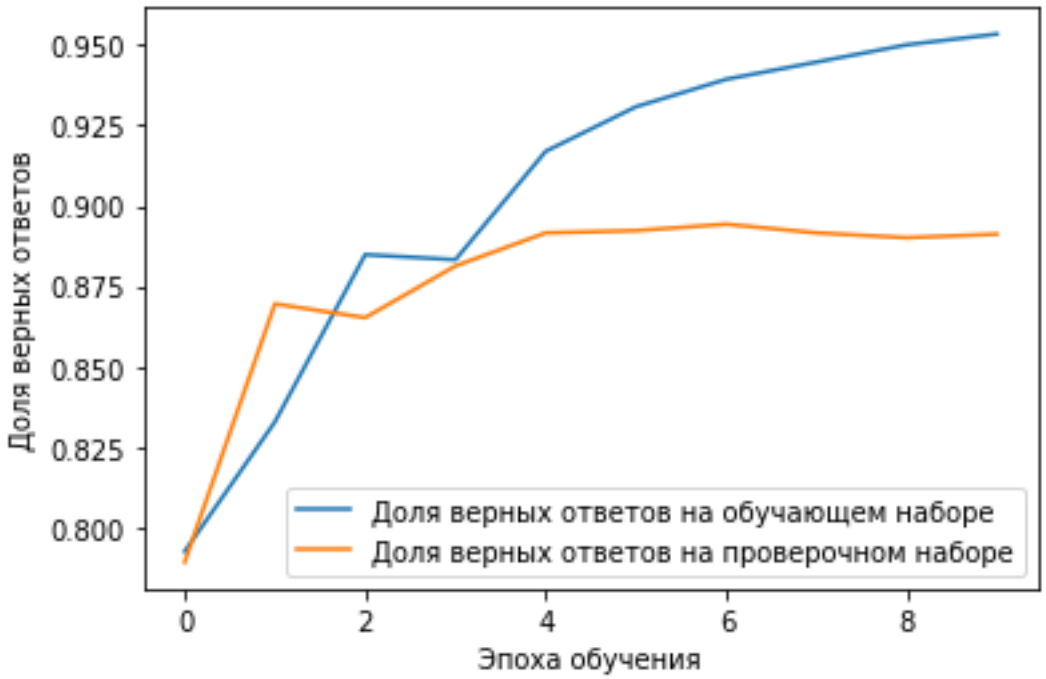

Рисунок 3.8. Обучение нейронной сети

Далее после обучения нейронной сети на основе обучающей выборки, она сможет предсказывать настроения отзыва просто пройдясь по нему. Для оценки точности модели, воспользуемся матрицей неточностей (Рисунок 3).

Чтобы правильно посчитать насколько модель точна, необходимо воспользоваться метриками точности и полноты.

Чтобы рассчитать значения точности и полноты, в начале строится матрица неточностей (ошибок) классификации (Таблица 3.2), сочетающая в себе все комбинации экспертной и системной оценки

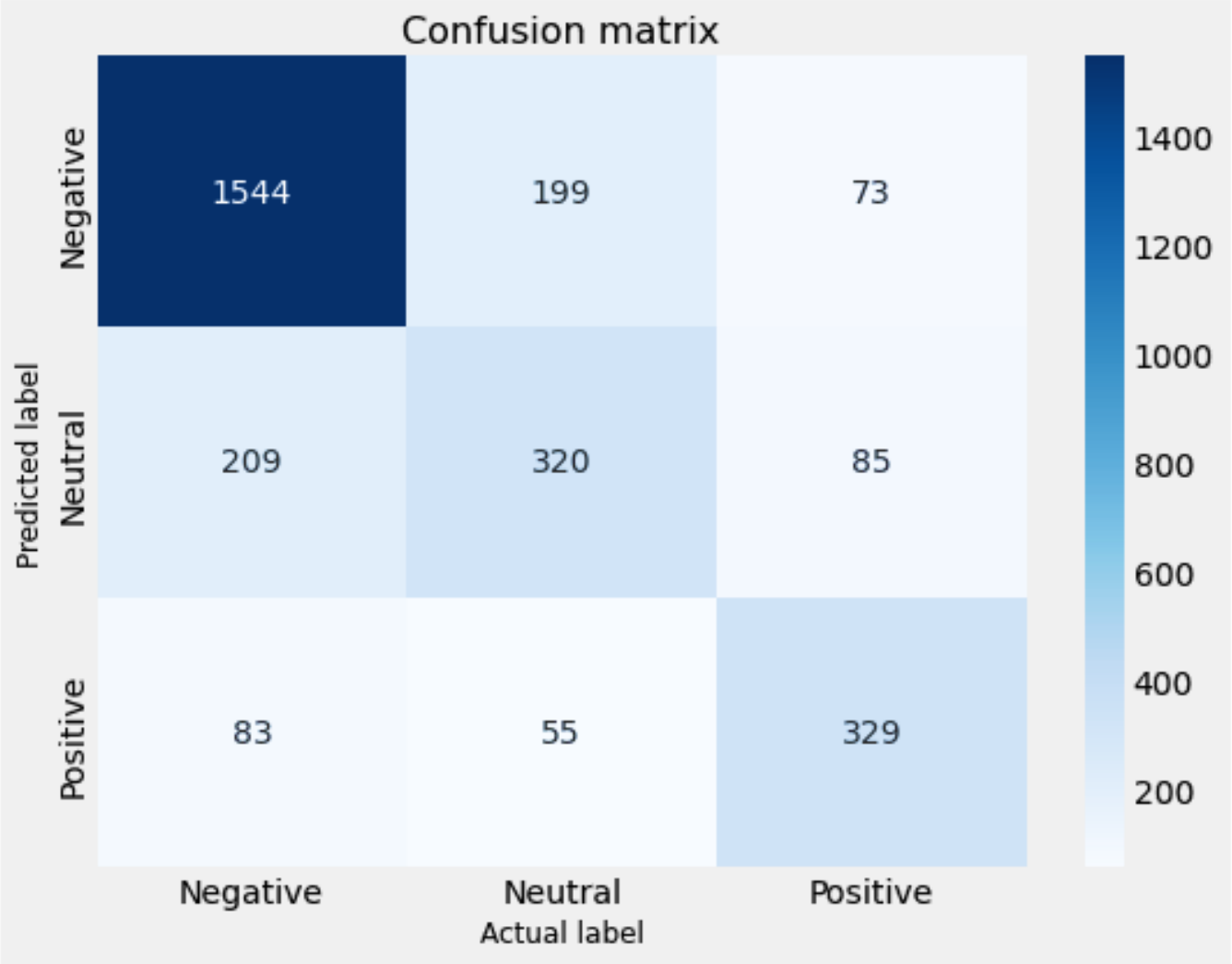

Рисунок 3. Матрица конфигуразии

Первая строка предназначена для отзывов, фактическое значение тональности которых в тестовом наборе равно 1. Из полученных данных видно, что из 10 000 отзывов значение тональности 1544 из них равно 1; и из этих 5000 классификатор правильно спрогнозировал 2643 из них как 1.

Это означает, что для 10 тысяч отзывов фактические значения тональности были равны 1 в тестовом наборе, и классификатор также правильно предсказал их как 1. Однако, в то время как фактические метки 5000 отзывов были равны 1, классификатор предсказал их как 0, что довольно неплохо.

Таким образом, он проделал хорошую работу в прогнозировании отзывов со значением тональности 0.

Матрица путаницы хороша тем, что она показывает способность модели правильно предсказывать или разделять классы. В конкретных случаях трочного классификатора, таких как этот пример, мы можем интерпретировать эти числа как количество истинных положительных, ложных, истинно отрицательных и ложно отрицательных результатов, а также смешанных моделей.

Таким образом можно рассчитать точность и полноту полученных данных:

$$F - \text{мера} = 0.8913078535720045$$

$$Recall = 0.8998697492673396$$

$$Precision = 0.8829073482428115 \#(3.2)$$

Эти значения является очень хорошими показателями для примера классификации данных, а также означает высокую долю корректно классифицированных примеров нейронной сетью, обученной на данных с отзывами «IMDb»

Матрица путаницы хороша тем, что она показывает способность модели правильно предсказывать или разделять классы.

В данной статье был рассмотрен набор данных, взятых с сайта IMDB, проведена обработка этих данных, их структуризация и очистка от частиц речи, после чего текст отзывов векторизовали, для того, чтобы было удобнее их обрабатывать методом Наивного Байеса, после параметров прогнозирования с результатами нейронной сети, можно сделать вывод о том, что хоть метод Наивного Байеса и проще и быстрее, но он не учитывает, всех параметров отзыва, поэтому применение нейронных сетей для классификации будет более подходящим решением.

Таким образом можно сделать вывод о целесообразности использования нейронной сети на большом объеме данных, при условии большой точности данного метода, досочного объема обучающей выборки, исключения проблемы переобучения модели.

Список литературы:

1 Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.

2 Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. — 2017. — Т. 30 №1. — С. 85-99

3 Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин). — 2007. — С. 141

4 Оценка классификатора (точность, полнота, F-мера) [Электронный ресурс] URL: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html/

5 Современные методы обработки естественного языка / Б. О. Близнюк [и др.] // Весник Харьковского национального университета имени В. Н. Каразина. Серия «Математическое моделирование. Информационные технологии. Автоматизированные системы управления». — 2017. — №16. — С. 14-26.

6 Шишкевич С.С. Аддитивная регуляризация наивного линейного байесовского классификатора. — Москва, 2016. — 27 с.