Абашева Ирина Владимировна,
Кандидат технических наук, Старший научный сотрудник –
старший бортовой оператор 442 лаборатории военного института
(научно-исследовательского), г. Санкт-Петербург

ПРОБЛЕМНЫЕ ВОПРОСЫ АНАЛИЗА ТЕКСТА СО СМЕШАННЫМ КОДОМ

Аннотация. В статье рассматривается проблема классового дисбаланса, которая является одним из важных вопросов оценивания эмоционального содержания сообщений со смешанным кодом. Предлагается подход к анализу эмоционального содержания для текстовых сообщений со смешанным кодом с использованием метода выборки в сочетании с метриками расстояния Левенштейна. Приводится сравнение характеристик различных подходов к машинному обучению.

Ключевые слова: текстовые сообщения со смешанным кодом, эмоциональное содержание, обучение машинное обучение, выборка обучающая выборка.

ВВЕДЕНИЕ

Систематизация особенностей поступающей телеметрической информации выявила ее неоднородность, что не позволяет классическими методами провести синтез алгоритмов оперативной обработки и выполнить требования по качеству обработки информации в условиях жестких временных ограничений [1]. Из-за быстрого развития социальных сетей пользователи могут легко обмениваться, обсуждать или передавать свою информацию, которая охватывает множество интересов, включая политику, обзоров различных продуктов, научных новинок и многое другое. В связи с локдауном COVID-19 количество онлайн-пользователей в социальных сетях значительно увеличилось. Кроме то-

го, сайты социальных сетей предоставляют пользователям возможность создавать контент на родном языке или на смешанном языковом коде [2]. Под смешанным кодом в статье подразумеваются тексты, включающие в себя разные языки и символьные конструкции для выражения своих взглядов при общении в социальных сетях.

Наличие данных со смешанным кодом значительно усложняют процедуры машинного перевода, поиска информации, идентификации языка, анализа настроений и другие задачи обработки текстовой информации. Анализ эмоционального состояния автора в текстах со смешанным кодом представляет собой сложную задачу, так как обычные методы предварительной обработки данных, такие как морфологический анализ, недостаточны. Это происходит потому, что данные со смешанным кодом не имеют строгой грамматической структуры и могут содержать скрытые лексические элементы.

Основными проблемами анализа эмоционального содержания текста, а именно извлечения настроений в смешанном социальном тексте со смешанным кодом, являются:

- отсутствие строгой грамматической структуры и порядка слов в предложениях со смешанным кодом;

- различные варианты написания слов, без правил орфографии, что затрудняет нормализацию при анализе;

- креативные варианты написания и орфографии, которые создают дополнительные сложности при анализе настроений;

- использование аббревиатур, которые нужно учитывать при анализе;

- отсутствие заглавных букв, что затрудняет идентификацию начала предложений.

С другой стороны, контроль эмоционального содержания сообщений позволит своевременно выявлять авторов текстов, находящихся в нестабильном эмоциональном состоянии, что позволит своевременно предотвращать различ-

ные чрезвычайные ситуации как в социальной сфере, так и в сфере государственной безопасности [3].

Для разрешения указанной проблемы требуется решить следующие частные задачи.

1. Классифицировать данные со смешанным кодом и текстовые данные, используя усовершенствованный алгоритм проверки орфографии.

2. Построить словарь лексики для этого смешанного корпуса слов и нормализовать слова орфографическими вариациями в корпусе с помощью метрики расстояния Левенштейна.

3. Применить различные методы машинного обучения для извлечения тональностей и решить проблему дисбаланса классов, используя методы выборки.

КЛАССИФИКАЦИЯ ЭМОЦИОНАЛЬНОГО СОДЕРЖАНИЯ ТЕКСТА

В интересах решения поставленных задач был проведен обзор литературы зарубежных авторов в области сентиментального анализа «sentimental analysis» одноязычных данных, сентиментальному анализу данных со смешанным кодом, а также в области дисбаланса классов при сентиментальном анализе [4, 5]. В качестве концептуальных основ анализа эмоционального содержания текста предлагается использовать элементы теории сентиментального анализа.

В начале XXI века анализ эмоционального содержания был сложной задачей в области обработки естественного языка. Первые работы в этой области были посвящены маркетингу, например, сентиментальный анализ использовался в исследованиях управления для выведения цены продукта, для автоматического определения мнения о продуктах, прогнозирования цен и получения обратной связи [6].

Анализ эмоционального содержания возможен на трех уровнях, а именно: уровень документа, уровень предложения и уровень сущности. Методы анализа на всех этих уровнях можно разделить на три категории, а именно: подходы, основанные на словаре, методы машинного обучения и гибридные подходы. Под-

ходы, основанные на словарях, в основном сосредоточены на предопределенных лексиконах [7]. Словарные подходы могут хорошо работать с одноязычными данными, поскольку стандартные лексиконы строятся и хранятся в словаре. Словарные подходы не требуют обучающих данных и крайне сложны для междоменных или многоязычных данных [8].

Представленные подходы не являются универсальными, так, например, для анализа эмоционального окраса одноязычных данных хорошо подходят методы машинного обучения, такие как контролируемый, неконтролируемый и полуконтролируемый подходы. Однако основным недостатком методов машинного обучения является необходимость в большом объеме обучающей выборки, специфичных для конкретной области.

Для решения проблемы дисбаланса классов были предложены различные гибридные методы машинного обучения, такие как ROSE, методы избыточной выборки, методы ресамплинга векторного пространства и т.д. Считается, что метод избыточной выборки решает проблему дисбаланса классов для данных со смешанным кодом социального текста. В настоящее время расстояние Левенштейна до сих пор не использовалось в качестве метода предварительной обработки для решения орфографических вариаций в смешанных данных. Кроме того, методы ресамплинга никогда не применялись для данных со смешанным кодом для решения проблемы дисбаланса классов.

ЭКСПЕРИМЕНТ

Для практической реализации предлагаемого в статье подхода эмоционального анализа текста был создан двуязычный набор данных из социального текста со смешанным кодом. Набор данных состоял из 15 744 предложений и был разделен на три категории: 11 335 предложений рассматриваются как обучающие данные, 1260 предложений рассматриваются как валидационный набор и 3149 предложений используются в качестве тестовых данных. Настроения, присутствующие в наборе данных, разделены на четыре категорий (таблица 1), а именно: положительные, отрицательные, смешанные чувства и неизвестное состояние.

Таблица 1

Описание набора данных

Категории	Обучающие данные	Данные проверки	Тестовые данные
Положительная	7627	856	2076
Отрицательная	1816	195	597
Смешанные чувства	1283	141	377
Неизвестное состояние	609	68	173

Предлагаемая архитектура данного метода показана на рисунке 1.

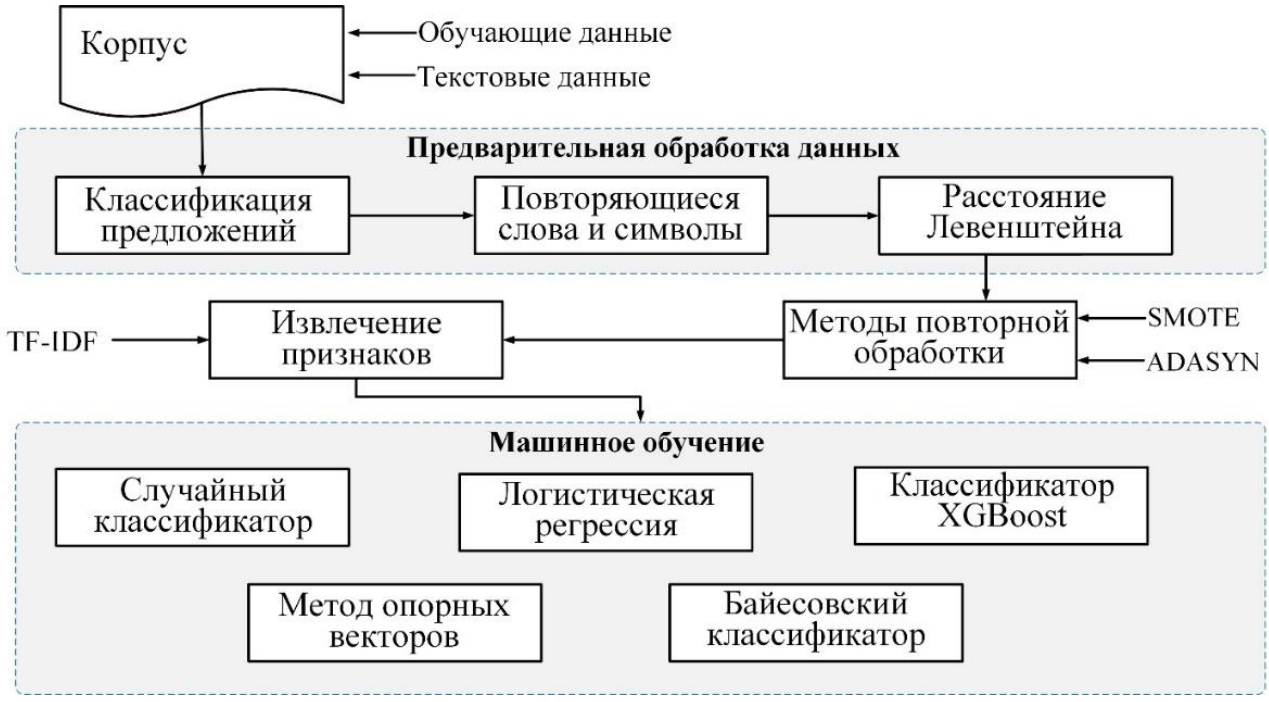

Рис.1. Архитектура предлагаемой методологии

Предварительная обработка текста является важным шагом, который позволяет преобразовать и извлечь значимую информацию из данных. Метод предварительной обработки (также называемый очисткой данных) устраняет ошибки и несоответствия, присутствующие в данных [9]. Иногда несоответствия данных нелогичные сообщения или скрывают важную информацию, что влияет на точность распознавания.

На первом этапе предварительной обработки необходимо произвести классификацию предложений и удалить специальные символы.

Алгоритм классификации предложений используется для предложений со смешанным кодом S_{sm} и несмешанных S_{nm} предложений. Все предложения $S = S_{sm} \cup S_{nm}$ даются в качестве входных данных для алгоритма и делятся на массивы слов $\{w_1, w_2, \ldots, w_n\}$. Для каждого слова в предложении проверяется, принадлежит ли оно основному языку или нет, с помощью детектора языка. Предложение, содержащее слова неосновного языка, считается предложениями со смешанным кодом.

Необходимо учитывать, что в социальных сетях люди, как правило, используют много вариантов одного слова, чтобы выразить свои разнообразные эмоции. Например, люди вводят слово «вау», чтобы выразить свою повышенную степень эмоций. Такой вариант слов должен быть нормализован до извлечения признаков, что необходимо для более точного захвата чувств. Повторяющийся символ должен выявляться путем сравнения его с базовой словоформой. Поскольку эксперимент нацелен на обнаружение положительных эмоций и не направлен на классификацию различных степеней в положительных эмоциях, повторяющиеся слова удаляются как часть процедуры нормализации.

На втором этапе предварительной обработки вычисляется расстояние Левенштейна. Расстояние Левенштейна выясняет минимальное количество односимвольных правок, которое требуется для нормализации его с базовым словом [10]. Расстояние Левенштейна имеет четыре операции, а именно: идентификация, вставка, подстановка и удаление.

Пусть $d(i,j)$ – расстояние между символами слова w_i и символами слова w_j. Например, w_1 = Талайвар и w_2 = Thalaivar – два слова, где w_1 – исходное слово, а w_2 — целевое (базовое слово). $w_1[1.... m]$ и $w_2[1.... n]$, где m и n – длина слов. Когда каждый символ в w_1 сравнивается с целевым словом w_2, символ 'h' добавляется в w_2 чтобы обеспечить соответствие w_1. Минимальное расстояние редактирования, необходимое для нормализации слова от источника до целевой строки, равно 1. Установлено максимальное расстояние редактирования как два, замечено, что когда расстояние редактирования больше 2, слово превраща-

ется в другое слово, не соответствующее значению базового слова. Эти процедуры нормализации помогли снизить вычислительную сложность.

$$d(0,0) = 0,\ \ d(i,j) = \text{мин} \begin{cases} d\,(i-1,j)\ +1 & \text{стирание} \\ d\,(i,j-1)\ +1 & \text{вставка} \\ d\,(i-1,j-i)\ +1 & \text{замещение} \end{cases}$$

Термин «Частота» и «Обратная частота» документа используется в методе преобразования текста в векторную форму Tf-Idf. Tf-Idf предложен Джонсом в 1972 году и полезен для поиска информации и процесса классификации текста. При этом рассчитываются статистические меры, позволяющие присвоить вес каждому слову в корпусе. Частота (Tf) определяется как количество раз, когда слово встречается в документе, тогда как обратная частота документа (Idf) увеличивает вес редко встречающихся слов, но уменьшает вес часто встречающихся неважных слов. Извлечение функций для анализа сентиментального содержания может быть выполнено с помощью Bag of Words (BoW), Word2vec, Global Vector Representation (GloVe). BoW придает больший вес часто встречающимся словам в документе. Word2Vec, GloVe являются предварительно обученной моделью для преобразования текста в вектор. Поскольку это смешанные данные с кодом, предварительно обученные встраивания слов недоступны. Кроме того, смешанные кодовые данные обычно содержат слова, которые имеют меньшую частоту, но нуждаются в весе, и, следовательно, Tf-Idf был выбран для извлечения функций [11].

После завершения методов предварительной обработки следующей задачей является применение алгоритмов машинного обучения для классификации тональности данных со смешанным кодом. Есть, по крайней мере, две проблемы, которые могут возникнуть при применении алгоритмов машинного обучения для любых данных. Первая задача заключается в том, чтобы найти подходящую технику извлечения признаков, а вторая – найти подходящий алгоритм классификации. При выборе наиболее подходящих методов извлече-

ния признаков любой алгоритм классификации может классифицировать настроения из данных, смешанных с кодом. Эксперимент проведен с использованием различных алгоритмов классификации, а именно: вероятностных моделей (наивный байесовский классификатор (NB)), линейного классификатора (классификатор опорных векторов машин (SVM)), основанных на решениях (random Forest Classifier и классификатор XGBoost) и статистической модели (логистическая регрессия).

Точность F1-Score и Accuracy являются обычно используемыми оценочными мерами для оценивания ошибок классификации, но при использовании алгоритмов машинного обучения для несбалансированного набора данных должна быть выбрана соответствующая метрика оценки [12]. Были проанализированы различные метрики оценки, такие как точность, балл F1 и макро-средний балл F1 (среднее значение балла F1). Поскольку точность и балл F1 не являются предпочтительными метриками для оценки данных о дисбалансе классов, макро-средний балл F1 выбирается вместе с отзывом и точностью в качестве оценочных метрик. F1 Score – это взвешенное среднее значение точности и запоминаемости. Чтобы подчеркнуть важность идентификации данных со смешанным кодом при анализе тональностей идентифицировали настроения со смешанными данными кода и без них.

В целях проведения анализа эмоциональности без разделения смешанных и некодовых смешанных текстов был подготовлен набор данных состоящий из 11 335 обучающих предложений и 3149 тестовых предложений. Использовались два метода ресамплинга, а именно методы SMOTE и ADASYN, и их производительность была сопоставлена. В таблице 2 представлен результат, полученный с использованием метода выборки SMOTE, а в таблице 3 приведены результаты, полученные с помощью метода ADSYN. Исходя из представленных результатов можно сделать вывод, что эти методы ресамплинга улучшили F1-балл на 50%.

Таблица 2

Значение F1-балла с использованием метода SMOTE

Модель	Положительная	Отрицательная	Смешанные Чувства	Неизвестное состояние
Классификатор случайных лесов	0.81	0.25	0.04	0
Логистическая регрессия	0.73	0.26	0.11	0
Классификатор XGBoost	0.8	0.28	0.02	0
Опорные вектора	0.8	0.22	0.08	0
Наивный Байес	0.9	0.3	0	0

Таблица 3

Значение F1-балла по методике ADASYN

Модель	Положительная	Отрицательная	Смешанные чувства	Неизвестное состояние
Классификатор случайных лесов	0.81	0.24	0.03	0.02
Логистическая регрессия	0.81	0.26	0.03	0
Классификатор XGBoost	0.8	0.26	0.02	0.02
Опорные вектора	0.8	0.21	0.08	0
Наивный Байес	0.89	0.21	0	0

Из таблиц 2 и 3 видно, что логистическая регрессия работает по сравнению с другими методами, тогда как наивный байес не может предсказать другие классы, кроме положительных. Кроме того, можно заметить, что даже после применения методов выборки алгоритмы классификации более смещены в сторону положительного класса. Этот дисбаланс классов обусловлен неразделением смешанных и некодовых смешанных данных [12].

В целях проведения анализа эмоциональности после разделения классификационных кодовых смешанных и некодовых смешанных данных была сформирована выборка из 15744 предложений, при этом 541 предложение представляет собой предложения без кода, которые идентифицируются и удаляются перед предварительной обработкой данных. Это привело к лучшему

распределению настроений по всем категориям, что очень хорошо видно из таблиц 4 и 5, где был получен повышенный балл F1 для каждого класса, а также привело к улучшению усредненного балла F1, рассчитанного для всех классов. Среднее значение F1-балла увеличилось на 2% после удаления данных, не связанных со смешанным кодом.

После удаления данных, не связанных со смешанным кодом, было отмечено, что методы выборки работают лучше даже для категории «Смешанные чувства» по сравнению с предыдущими таблицами 1 и 2 из-за использования методов ресамплинга SMOTE и ADASYN. Но было замечено, что классовый дисбаланс все еще существует, несмотря на все эти усилия.

Таблица 4

Значение F1-балла с использованием метода SMOTE

Модель	Положительная	Отрицательная	Смешанные Чувства	Неизвестное состояние
Классификатор случайных лесов	0.7	0.34	0.24	0.12
Логистическая регрессия	0.68	0.34	0.29	0.26
Классификатор XGBoost	0.79	0.27	0.2	0.26
Опорные вектора	0.8	0.17	0.22	0.19
Наивный Байес	0.8	0.4	0	0

Таблица 5

Значение F1-балла по методике ADASYN

Модель	Положительная	Отрицательная	Смешанные чувства	Неизвестное состояние
Классификатор случайных лесов	0.81	0.28	0.24	0.08
Логистическая регрессия	0.81	0.28	0.23	0.12
Классификатор XGBoost	0.8	0.29	0.23	0.12
Опорные вектора	0.8	0.22	0.21	0.1
Наивный Байес	0.78	0.3	0	0

До сих пор все эксперименты проводились без учета этапа предварительной обработки расстояния Левенштейна. Стоит отметить, что расстояние Левенштейна играет важную роль в идентификации данных, смешанных с кодом. Как известно,

расстояние Левенштейна помогает минимизировать орфографические ошибки, используя минимальное расстояние редактирования. В данных со смешанным кодом орфографические вариации являются серьезной проблемой, но могут быть смягчены с помощью расстояния Левенштейна. В таблицах 6 и 7 рассматривается результат после применения расстояния Левенштейна.

Таблица 6

Значение F1-балла с использованием метода SMOTE

Модель	Положительная	Отрицательная	Смешанные чувства	Неизвестное состояние
Классификатор случайных лесов	0.8	0.44	0.34	0.18
Логистическая регрессия	0.68	0.37	0.19	0.16
Классификатор XGBoost	0.8	0.44	0.21	0.22
Опорные вектора	0.8	0.43	0.24	0.18
Наивный Байес	0.8	0.3	0	0

Таблица 7

Значение F1-балла по методике ADASYN

Модель	Положительная	Отрицательная	Смешанные чувства	Неизвестное состояние
Классификатор случайных лесов	0.81	0.36	0.22	0.06
Логистическая регрессия	0.81	0.39	0.12	0.12
Классификатор XGBoost	0.8	0.29	0.13	0.21
Опорные вектора	0.8	0.34	0.21	0.2
Наивный Байес	0.78	0.32	0	0

Общая производительность почти аналогична предыдущим результатам, показанным в таблице 6. Поскольку корпус был построен путем извлечения комментариев YouTube, в них много несоответствий, таких как орфографические ошибки, многие слова, смешивающие код языка и т. д.

РЕЗУЛЬТАТЫ

На рисунке 2 показана производительность каждого отношения с использованием различных подходов машинного обучения. Расстояние Левенштейна пре-

одолело орфографические ошибки, присутствующие в данных, и многие сильные особенности для каждого класса в некоторой степени облегчили проблему классового дисбаланса. Кроме того, было отмечено, что один алгоритм машинного обучения работает лучше для конкретного класса и не работает хорошо при рассмотрении всех классов. Это было основной причиной попыток многих классификаторов машинного обучения проверить их многоклассовую обработку.

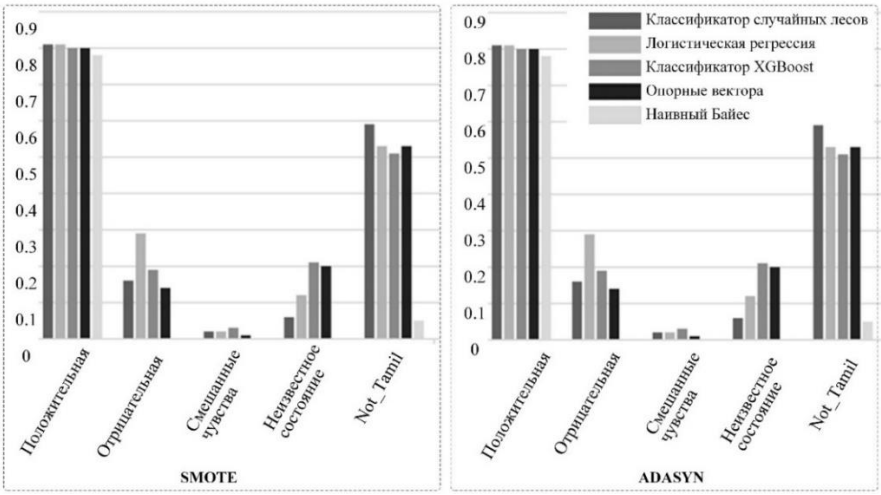

Рис. 2. Результаты всех отношений с использованием подходов машинного обучения

Из всех методов машинного обучения логистическая регрессия показала лучшую производительность по сравнению с остальными классификаторами.

ЗАКЛЮЧЕНИЕ

В статье рассмотрен подход к анализу эмоционального содержания текста со смешанным кодом. Предложено в качестве метода предварительной обработки смешанных данных использовать расстояние Левенштейна, что улучшило результаты, поскольку оно хорошо работало с выявлением вариантов орфографии, которые сохранялись в смешанных данных, написанных в социальных сетях. Полученные данные показали, что проблему дисбаланса классов можно облегчить, удалив смешанные данные, не относящиеся к коду. Сочетание расстояния Левенштейна с методами выборки помогло увеличить F1-Score.

Дальнейшая работа может быть направлена на улучшение оценки F1 путем поиска сильных методов извлечения признаков или гибридных подходов, которые могут помочь решить проблему дисбаланса класса, существующую в смешанных с кодом социальных текстовых данных.

Список литературы:

1. Асиф М., Иштиак А., Ахмад Х., Альджуаид Х., Шах Дж.: Анализ настроений экстремизма в социальных сетях на основе текстовой информации. Телематика Информ. 48, 101345 (2020) (https://doi.org/10.1016/j.tele.2020.101345 – дата обращения 02.02.2023).

2. Денисова Е.А. Смешение языковых кодов в англоязычной литературе. Ленинградский государственный университет имени А.С. Пушкина С.49-57 (https://pdc-journal.com – дата обращения 04.02.2023).

3. Насукава, Т., Йи, Д.: Анализ настроений: захват благоприятности с использованием обработки естественного языка. В: Труды 2-й Международной конференции по сбору знаний, стр. 70–77 (2003). (https://doi.org/10.1145/945645.945658 – дата обращения 02.02.2023).

4. Лопес, В., Фернандес, А., Гарсия, С., Паладе, В., Эррера, Ф.: Понимание классификации с несбалансированными данными: эмпирические результаты и современные тенденции использования внутренних характеристик данных. Inf. Sci. 250, С.113–141 (2013) (https://doi.org/10.1016/j.ins.2013.07.007 – дата обращения 02.02.2023).

5. Lu, Y., Cheung, Y., Tang, Y.: Индекс влияния дисбаланса Байеса: мера несбалансированного набора данных класса для задачи классификации. IEEE Trans. Нейронная сеть. Учиться. 31(9), С.3525–3539 (2019).

6. Chen, J.I.Z., Lai, K.-L.: Управление энергией на основе машинного обучения на узлах сети Интернета вещей. J. Расчет тенденций. Sci. Smart Technol. 3 (2020), 127–133 (2020).

7. Якобсон Р. О лингвистических аспектах перевода. Вопросы теории перевода в зарубежной лингвистике. Москав, 1978. С.16-24.

8. Дурайпандян М.: Оценка производительности алгоритма маршрутизации для Мане на основе методов машинного обучения. J. Расчет тенденций. Sci. Smart Technol. (TCSST) 1(01), С.25–38 (2019).

9. Нгуен, Т., Нгуен, Л., Цао, Т. (2017). Анализ тональности медицинского текста с использованием комбинации машинного обучения и оценки SO-CAL. В: Азиатско-Тихоокеанский симпозиум по интеллектуальным и эволюционным системам (IES), С.49-56. (https://doi.org/10.1109/IESYS.2017.8233560 – дата обращения 02.02.2023).

10. Левенштейн, В.И.: Двоичные коды, способные исправлять удаления, вставки и развороты. Сов. Доктор философии 10, 707с. (1966).

11. Haixiang, G., Li, Y., Shang, J., Mingyun, G., Yuanyue, H., Gong, B.: Обучение на несбалансированных по классам данных: обзор методов и приложений. Expert Syst. Appl. (2016) (https://doi.org/10.1016/j.eswa.2016.12.035 – дата обращения 02.02.2023).