Анализ точности моделей машинного обучения с использованием методов векторизации для задач классификации разнородных текстовых данных

А.Н. Алпатов, К.С. Попов, А.Н. Чесалин

***Аннотация* — В данной работе исследуется задача обработки естественного языка с использованием методов машинного обучения, в частности, классификации неструктурированных разнородных текстовых наборов данных. В статье представлен сравнительный анализ некоторых актуальных и широко используемых методов и моделей машинного обучения с учителем, применяемых для мультиклассовой классификации на разнородных текстовых источниках данных с использованием различных методов извлечения признаков. Рассматривается зависимость точности предсказания классов моделями классификаторов от качества использованных в данной работе корпусов текстовых данных, применяя различные методы векторизации на обработанном наборе исходных данных. На основе проведенного анализа, предложена обобщенный схема функционирования программного обеспечения, реализующая алгоритм построения модели классификации неструктурированных текстов, в виде конвейера по обработке текстовых корпусов и управлению моделями машинного обучения. В ходе проведенного эксперимента продемонстрировано, что для корпусов с различным качеством исходных текстовых данных, точность предсказаний классификаторов разнится. Данное обстоятельство проявилось в том, что классификаторы обладают более низкими характеристиками на корпусе текстов музыкальных композиций и высокие на текстах новостных сводках. При этом показано, что при определённых условиях, использование решений для повышения качества классификации, таких как стекинг и добавление дополнительных признаков классификации, может приводить не к улучшению, а, наоборот, к ухудшению результатов предсказания классов, что, в конечном итоге, может негативно сказаться на конечной точности получаемых результатов модели.**

***Ключевые слова* — классификация, обработка естественного языка, машинное обучение, текстовый корпус, векторизация, векторные модели, числовые метрики, genism, scikit-learn, TF-IDF, Word2Vec, Doc2Vec.**

Статья получена 20 апреля 2022.

Алексей Николаевич Алпатов к.т.н. доцент кафедры инструментального и прикладного программного обеспечения, ИИТ, РТУ МИРЭА, Москва, Россия (e-mail: alpatov@mirea.ru).

Кирилл Сергеевич Попов, студент РТУ МИРЭА (e-mail: popov.k.s2@edu.mirea.ru).

Александр Николаевич Чесалин к.т.н., доцент кафедры компьютерной и информационной безопасности, ИИИ, РТУ МИРЭА, Москва, Россия (e-mail: chesalin@mirea.ru).

I. Введение

Классификация текстовых данных является одной из распространенных задач, решаемых в рамках задачи обработки естественного языка (англ. NLP — Natural language processing). Использование методов машинного обучения для классификации текстовых данных получило широкое распространение в различных областях применения, таких как: классификация новостей и документов, фильтрация спама, определение эмоциональной окраски отзыва, а также для других прикладных задач [1].

Актуальность данной задачи определяется стремительным ростом количества накапливаемой и обрабатываемой информации, при этом использование классификаторов позволит отсортировать, ограничить поиск и тем самым ускорить получение необходимой информации, реорганизовав ее по подмножествам [2].

С целью ускорения обработки естественного языка, в данной статье произведён критический анализ моделей классификации, а также исследуются подходы к предобработке и векторизации текстов и анализируется точность моделей классификации на двух разнородных источниках текстовых данных.

Для обработки неструктурированных корпусов текстовых данных в работе предлагается использование разработанного программного обеспечения, обобщённая структурно-функциональная схема, которого представлена на рис.1.

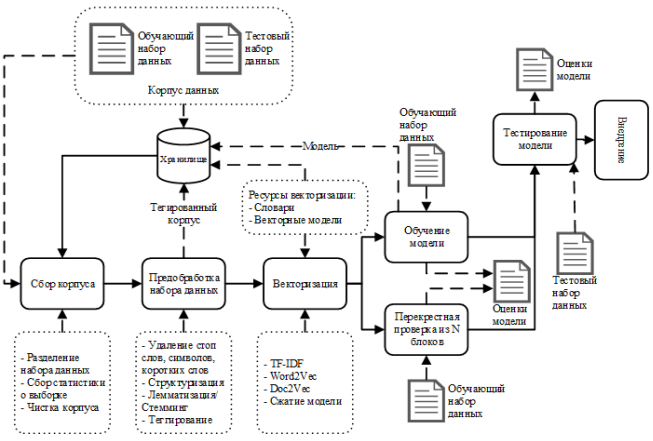


Рис. 1. Обобщенная схема программного обеспечения

для обработки неструктурированных корпусов текстовых данных

II. Предварительные сведения

A. Анализ внутренней структуры источников данных

При построении моделей классификации первым и наиболее важным этапом является подбор и создание текстовых данных, пригодных для обучения модели, так называемых корпусов. Под корпусом в данной работе будем понимать совокупность взаимосвязанных по какому-либо принципу документов или текстов на естественном языке. Корпусы могут быть как аннотированными, т.е. текстовые данные снабжаются специальными метками классов для алгоритмов моделей обучения с учителем, или неаннотированными для тематического моделирования и кластеризацию, применяемые в моделях обучения без учителя [4].

Для данной работе, с целью проведения эксперимента, выбраны два разнородных корпуса данных с различным качеством текста для анализа.

В качестве первого корпуса для анализа был взят готовый источник данных из 300 000 текстов музыкальных композиций, разделенных на 10 жанров [5]. Набор данных представлен в виде csv файлов, разделенных на две части – тестовую и обучающую выборку, и содержат в себе информацию о наименовании композиции, исполнителе, годе выпуска, альбома, языке и тексте каждой композиции.

В рамках задачи классификации на данном наборе данных наибольший интерес представляет определение жанровой принадлежности текста композиции.

Для дальнейшей работы с корпусом необходимо произвести его «очистку» – удалить все композиции не на английском языке, некорректные данные, дубликаты текстов, инструментальные композиции, и произведения записанные в виде табулатуры. В результате процесса «очистки» корпус стал содержать данные, представленные в таблице I.

Таблица I – Данные очищенного корпуса с текстами композиций

Жанр	Исполнители		Композиции	
	Количество	Процент (%)	Количество	Процент (%)
Country	246	2.325	1864	0.864
Electronic	467	4.414	1973	0.915
Folk	449	4.244	7609	3.528
Hip-Hop	470	4.442	2179	1.01
Indie	767	7.249	6659	3.087
Jazz	683	6.455	7923	3.673
Metal	1070	10.112	17650	8.183
Pop	2838	26.822	71071	32.948
R&B	222	2.098	2661	1.234
Rock	3369	31.84	96115	44.559
Всего	11689		215704	

Данные из таблицы демонстрируют неравномерное количества данных по классам, поэтому, в рамках данного исследования, было принято решение выбрать случайным образом 1800 композиций для каждого жанра, получив 18 000 записей для обучения моделей, и тем самым уравновесив вклад тренировочных данных для предсказания каждого класса.

На рис.2 представлено «облако тегов (слов)» из наиболее значимых 200 слов, вычисленных при помощи метода TF-IDF, в каждом из представленных жанров.

Рис.2. «Облака тегов» для каждого из музыкальных жанров датасета

Второй корпус для анализа представлен в виде набора новостей, разделенных на 10 групп, представленного в виде равного количества текстовых документов [6]. Для удобства работы с данном корпусом, содержимое текстовых файлов были перенесено в табличное представление в виде csv файла и каждой записи из набора была добавлена метка соответствующей группы.

Содержимое корпуса по новостным сводкам представлено в таблице II.

Таблица II – Содержимое корпуса по новостным сводкам

Группа новостей	Количество документов
business	100
entertainment	100
food	100
graphics	100
historical	100
medical	100
politics	100
space	100
sport	100
technologie	100
Всего	1000

На рис.3 представлены «облака тегов» из 200 наиболее часто встречаемых по группам новостей слов, вычисленных при помощи алгоритма TF-IDF.

Рис.3. «Облака тегов» для каждой из групп новостей

B. Предобработка корпусов данных

В качестве предобработки корпусов текстовых данных, для каждого из датасета, были выполнены следующие действия:

- Токенизация – документы были разбиты на мелкие структурные элементы, сначала на предложения затем на наименьшие символьные структурные единицы текста – токены, которые включают в себя слова, цифры, знаки.
- Чистка – корпус был очищен от стоп слов, знаков препинания и слов короче 3-х символов, не несущих смысловой нагрузки для последующих этапов анализа.
- Нормализация – очищенный массив токенов был приведен к нормализированному виду – токены приведены к нижнему регистру и снабжены тегами частей речи.
- Лемматизация – каждый токен был приведен к смысловой канонической форме слова по указанному тегу части речи, называемой леммой.

Полученные наборы данных, сформированные из выбранных датасетов, после вышеописанных шагов предобработки могут обрабатываться с помощью методов машинного обучения для решения задач классификации.

Для задач подготовки и обработки текстовых данных использовался следующий стек модулей Python: pandas – для чтения, записи и организации выборки данных, представленных в табличном представлении, nltk – для организации функционала чтения корпуса, токенизации, нормализации и лемматизации, pickle – для промежуточного сохранения обработанного корпуса и его загрузки.

C. Модели векторизации

Для обучения моделей машинного обучения необходимо преобразовать текстовые данные в их векторное представление. Данный процесс на практике называется извлечением признаков или векторизацией [4].

Для подстановки текстового набора данных в векторном пространстве существует множество алгоритмов и моделей [4, 7-9].

В рамках данной работы был использован следующий перечень одних из самых популярных на практике моделей для векторизации документов из библиотек scikit-learn и gensim:

- TF-IDF [4,8,9];
- Hashing Vectorizer [10];
- Word2Vec [8,9];
- Doc2Vec [4,8].

Для использования моделей Word2Vec и Doc2Vec на обоих текстовых корпусах были обучены контекстно зависимые модели представления признаков в виде векторов на основе тренировочных наборов данных.

D. Модели классификаторов

В рамках эксперимента данной работы были выбраны следующие модели классификаторов [3]:

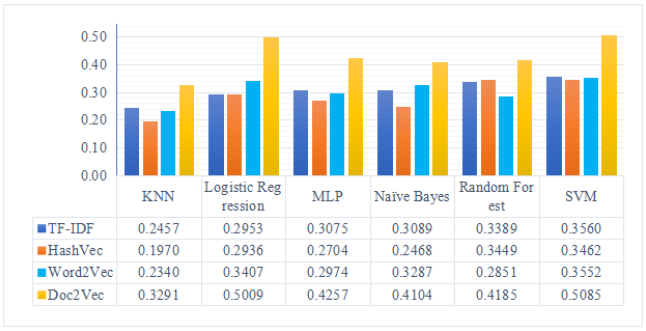
- Метод k-ближайших соседей (англ. k-nearest neighbors algorithm, KNN);
- Логистическая регрессия (англ. Logistic Regression);
- Метод опорных векторов (англ. Support Vector Machine, SVM);
- Случайный лес (англ. Random Forest);
- Наивный байесовский классификатор (англ. Naive Bayes, NB);
- Многослойный персептрон (англ. Multilayer Perceptron, MLP).

III. Результаты

A. Оценка моделей

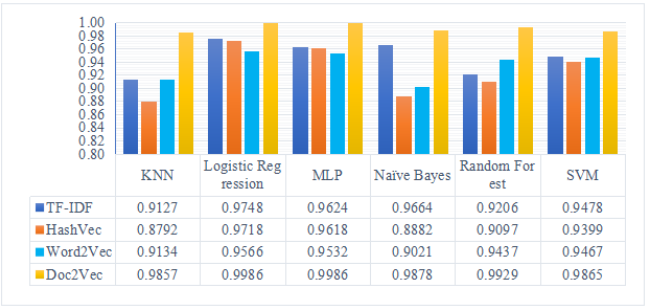
В рамках исследования к каждому из 6 рассматриваемых классификаторов, были применены 4 алгоритма векторизации документов. И на основании метода кросс-валидации на пяти блоках получены численные метрики для оценки точности классификации каждого из двух корпусов данных. В качестве метрик оценки точности предсказания моделей классификации использовались: точность (англ. precision, pre), полнота (англ. recall, rec) и f1-мера (англ. f1-score).

Усредненная численная метрика F1 моделей классификации для корпусов с текстами композиций и сводки новостей представлены в виде столбчатых диаграмм на рис.4 и рис.5, соответственно.



	KNN	Logistic Regression	MLP	Naïve Bayes	Random Forest	SVM
TF-IDF	0.2457	0.2953	0.3075	0.3089	0.3389	0.3560
HashVec	0.1970	0.2936	0.2704	0.2468	0.3449	0.3462
Word2Vec	0.2340	0.3407	0.2974	0.3287	0.2851	0.3552
Doc2Vec	0.3291	0.5009	0.4257	0.4104	0.4185	0.5085

Рис.4. Усредненные численные метрики моделей классификации для корпусов с текстами композиций



	KNN	Logistic Regression	MLP	Naïve Bayes	Random Forest	SVM
TF-IDF	0.9127	0.9748	0.9624	0.9664	0.9206	0.9478
HashVec	0.8792	0.9718	0.9618	0.8882	0.9097	0.9399
Word2Vec	0.9134	0.9566	0.9532	0.9021	0.9437	0.9467
Doc2Vec	0.9857	0.9986	0.9986	0.9878	0.9929	0.9865

Рис.5. Усредненные численные метрики моделей классификации для корпуса со сводкой новостей

Исходя из полученных результатов на кросс-валидации, можно сделать вывод, что два корпуса предоставляют различное качество исходных текстовых данных, что отражается на точности предсказаний классификаторов, как видно из представленных диаграмм, классификаторы показывают низкие оценки на корпусе текстов композиций и высокие на текстах новостных сводках.

Также можно заметить, что наилучшие результаты модели достигали с использованием метода индексации документов Doc2Vec, результативность же других методов разнится от использованной модели классификатора.

B. Тестирование моделей

Для тестирования были взяты обученные модели SVM совместно с векторной моделью Doc2Vec, так как данный стек методов показал одни из лучших результатов на тестовой выборке обоих корпусов.

При помощи инструментов из библиотеки yellowbrick были построены отчет о классификации и матрица смежности для тестового набора корпуса с текстами композиций, представленные на рис.6.

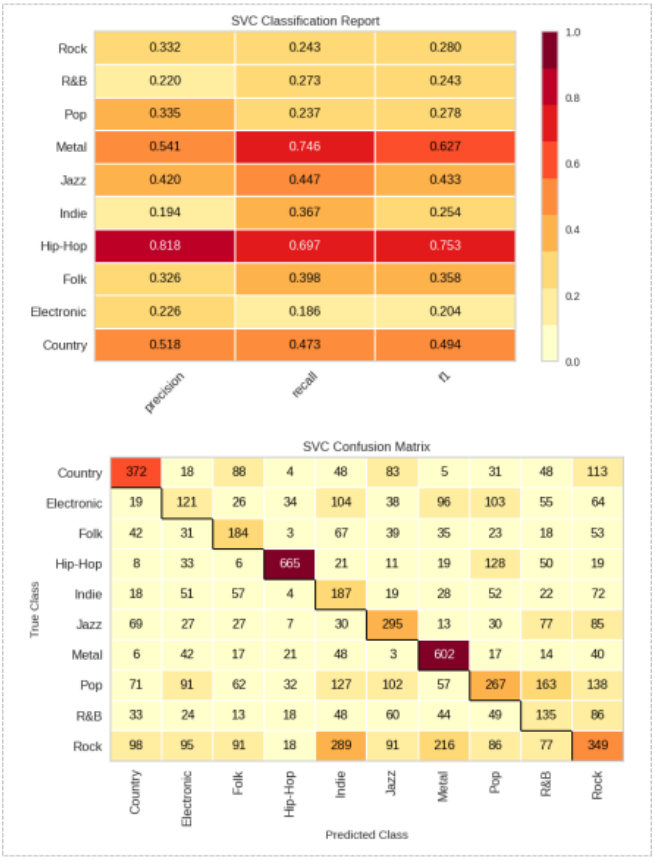


Рис.6. Отчет о классификации и матрица смежности для тестового набора корпуса с текстами композиций

Результаты показывают, как классификатору довольно проблематично отличить схожие жанры между собой владея лишь информацией о тексте композиций. Так, жанр «Rock» тесно переплетается с другими музыкальными направлениями, в частности «Indie» и «Metal», а тексты «Hip-Hop» частично определяются как «Pop», что может говорить о многообразии смыслового наполнения композиций и формировании музыкальных поджанров у каждого из классов. С другой стороны, такие результаты могут трактоваться за счет малой выборки, а также неровной и низкокачественной комплектацией корпуса. Наиболее хорошо распознанными классами на выборке стали «Hip-Hop» и «Metal», что говорит об их отличительном уникальном наполнении текстов.

Данный факт также может прослеживаться на диаграммах, построенных при помощи алгоритмов t-SNE и k-means, показанных на рис.7.

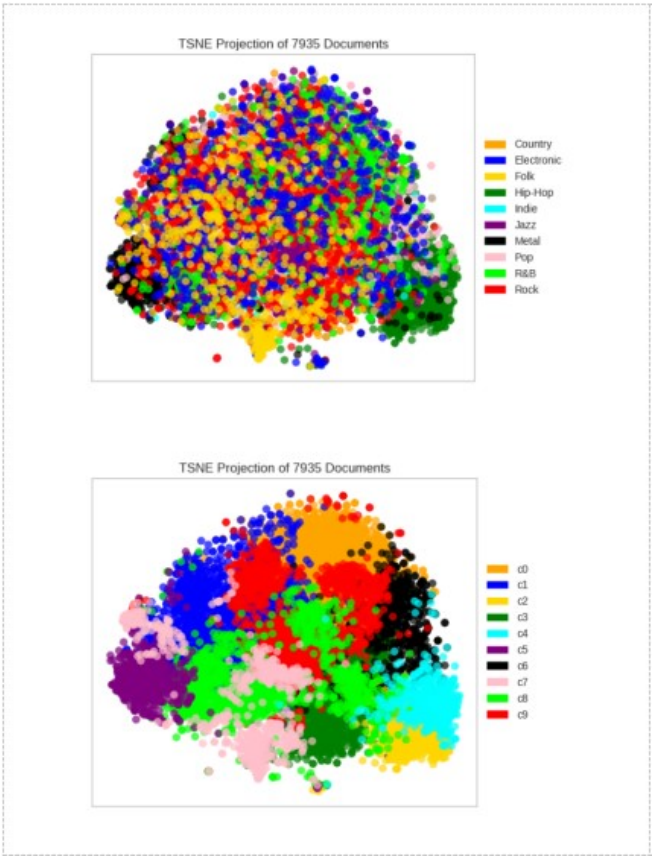


Рис.7. Визуализация данных корпуса музыкальных композиций при помощи алгоритмов t-SNE и k-means

Диаграммы, построенные этими двумя алгоритмами сильно схожи между собой, из правой части можно заметить, что визуализация документов классов «Hip-Hop» и «Metal» образуют отдельные большие скопления точек – кластеры, что также прослеживается и на правой диаграмме. Тексты других жанров в большей мере представляют хаотичный разброс точек в пространстве.

Стоит отметить, что использование существующих решений повышения качества классификации, такие как стекинг и добавление дополнительных признаков классификации (количество слов на документ, количество символов в документе и т.д.) привели к ухудшению результатов предсказания классов.

Для проверки работоспособности SVM на корпусе по сводкам новостей, данные были разделены в соотношении 70%/30% для обучения и тестирования соответственно. Отчет о классификации и матрица смежности для тестового набора корпуса с новостями, представлены на рис.8.

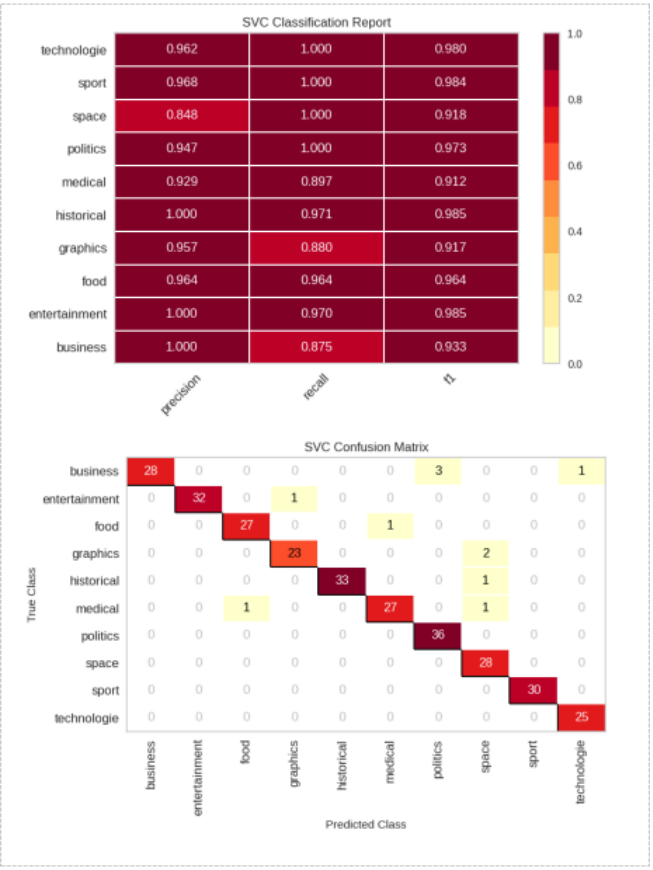


Рис.8. Отчет о классификации и матрица смежности для тестового набора корпуса с новостями

Из полученных результатов для новостного корпуса видно, что модель справилась с присвоением меток документам, несмотря на небольшую выборку данных для обучения.

Построенная диаграмма алгоритмом t-SNE на рис.9. Рис.9 визуализирует данные тестовой выборки в пространстве, образуя четкие кластеры для каждого класса документов.

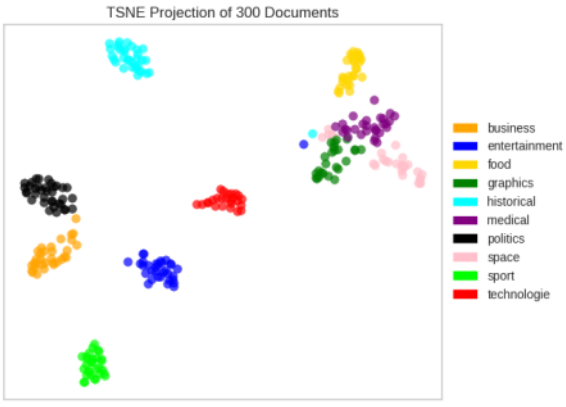


Рис.9. Визуализация данных при помощи t-SNE для новостного корпуса

IV. Заключение

В результате проведенного исследования методов классификации на примере двух разнородных корпусов текстов, получены следующие результаты:

- Наибольшая точность предсказания класса, вне зависимости от алгоритма классификации и степени

классифицируемости текстов, достигнута при использовании алгоритма векторизации Doc2Vec.

- Наибольшая точность классификации достигнута с применением алгоритмов SVM, Logistic Regression и MLP. Высокая точность при использовании данных моделей достигается только в совокупности с методом Doc2Vec, при использовании других методов векторизации, качество классификации значительно падает. При этом, отмечается, что использование других моделей также возможно, так как полученные результаты близки между собой и главным фактором в обеспечении точности является правильный выбор метода векторизации.
- Свойства используемых датасетов оказывают влияние на точности предсказаний классификаторов, что объясняется, прежде всего, объемом исходных текстовых данных. Но в то же время, использование существующих решений повышения качества классификации, такие как стекинг и добавление дополнительных признаков классификации (количество слов на документ, количество символов в документе и т.д.) для данных условий и используемых датасетов, привели не к улучшению, а, наоборот, к ухудшению результатов предсказания классов.
- Предложен обобщенный алгоритм построения модели классификации разнородных текстов, представленный в виде конвейера по обработке текстовых корпусов и управлению моделями машинного обучения.

Библиография

[1] G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] Епрев А.С. Автоматическая классификация текстовых документов. Математические структуры и моделирование. 2010. вып. 21, С.65-81.

[3] Полетаева Н.Г. Классификация систем машинного обучения Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. 2020. №1. С. 5-22.

[4] Федюшкин Н. А., Федосин С. А. О выборе методов векторизации текстовой информации. Научно-технический вестник Поволжья. 2019. Т. 6. С. 129-134.

[5] Multi-Lingual Lyrics for Genre Classification https://www.kaggle.com/datasets/mateibejan/multilingual-lyrics-for-genre-classification Дата обращения: 21.02.2022

[6] (10)Dataset Text Document Classification. https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification. Дата обращения: 21.02.2022

[7] Климов Д.В. Предобработка текстовых сообщений для метрического классифкатора. Символ науки. 2017. №12. С.25-32

[8] Мусаев А. А. и др. Обзор современных технологий извлечения знаний из текстовых сообщений. Компьютерные исследования и моделирование. 2021 Т. 13. № 6. С. 1291–1315 DOI: 10.20537/2076-7633-2021-13-6-1291-1315

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

[10] sklearn.feature_extraction.text.HashingVectorizer, scikit-learn 1.0.2 documentation https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.HashingVectorizer.html. Дата обращения: 3.04.2022

Accuracy analysis of machine learning models using vectorization methods for heterogeneous text data classification tasks

A.N. Alpatov, K.S. Popov, A.N. Chesalin

Abstract — This paper investigates the problem of natural language processing using machine learning techniques, in particular, classification of unstructured heterogeneous text data sets. The paper presents a comparative analysis of some relevant and widely used methods and teacher-assisted machine learning models used for multi-class classification on heterogeneous textual data sources using different feature extraction methods. The dependence of the accuracy of class prediction by classifier models on the quality of the text data corpora used in this paper, applying different vectorization methods on the processed set of source data, is considered. Based on this analysis, a generalized scheme of the software functioning, which implements the algorithm for constructing a model of classification of unstructured texts, in the form of a pipeline for processing text corpus and control of machine learning models is proposed. During the experiment, it was demonstrated that for corpora with different quality of initial text data, the accuracy of classifier predictions differed. This circumstance manifested itself in the fact that the classifiers have lower performance on the corpus of texts of musical compositions and high on the texts of news summaries. It is shown that under certain conditions, the use of solutions to improve the quality of classification, such as stacking and adding additional features of classification, can lead not to improvement, but on the contrary to the deterioration of the results of class prediction, which, ultimately, can have a negative impact on the final accuracy of the obtained model results.

***Keywords* — classification, natural language processing, machine learning, text corpus, vectorization, vector models, numerical metrics, genism, scikit-learn, TF-IDF, Word2Vec, Doc2Vec.**

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] Eprev A.S. Automatic classification of text documents. Mathematical structures and modeling. 2010. issue. 21, pp. 65-81.

[3] Poletaeva N.G. Classification of machine learning systems Bulletin of the Baltic Federal University. I. Kant. Series: Physical, mathematical and technical sciences. 2020. №1. pp. 5-22.

[4] Fedyushkin N. A., Fedosin S. A. On the choice of methods for vectorization of textual information. Scientific and technical bulletin of the Volga region. 2019. V. 6. pp. 129-134.

[5] Multi-Lingual Lyrics for Genre Classification [Online]. Available: https://www.kaggle.com/datasets/mateibejan/multilingual-lyrics-for-genre-classification. Accessed: 21.02.2022

[6] (10)Dataset Text Document Classification. [Online]. Available: https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification. Accessed: 21.02.2022

[7] Klimov D.V. Preprocessing of text messages for the metric classifier. Science symbol. 2017. No. 12. pp.25-32

[8] Musaev A. A. et al. Review of modern technologies for extracting knowledge from text messages. Computer research and modeling. 2021 Vol. 13. No. 6. pp. 1291–1315 DOI: 10.20537/2076-7633-2021-13-6-1291-1315

[9] Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. Automatic processing of texts in natural language and data analysis: textbook. allowance. Moscow.: Publishing House of the National Research University Higher School of Economics. 2017. 269 p.

[10] sklearn.feature_extraction.text.HashingVectorizer, scikit-learn 1.0.2 documentation [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.HashingVectorizer.html. Accessed: 3.04.2022