Классификация товаров по ОКПД2 кодам

Фирсов Сергей Андреевич
Кафедра интеллектуальных систем
МФТИ
firsov.sa@phystech.edu

Вознюк Анастасия Евгеньевна
Кафедра интеллектуальных систем
МФТИ
vozniuk.ae@phystech.edu

Старожилец Всеволод Михайлович
Кафедра интеллектуальных систем
Антирутина
vsevolod.starozhilets@antirutina.net

Аннотация

Исследование посвящено классификации товаров по кодам Общероссийского классификатора продукции по видам экономической деятельности (ОКПД 2), используя краткие текстовые описания. Основной целью является повышение точности и оптимизация ресурсов в процессе классификации за счёт анализа глубины кодов ОКПД2. Работа предлагает метод построения текстовых эмбеддингов, который способствует улучшению классификационных процессов в сфере логистики и учёта. В работе рассмотрены улучшения этого метода, основанные на иерархической структуре входных данных.

Ключевые слова ОКПД 2 коды · многоклассовая классификаци · иерархическая классификация · эмбеддинги · логистическая регрессия

1 Введение

Целью данного исследования является разработка и апробация метода классификации товаров по кодам ОКПД 2[1], используя краткие текстовые описания. Основная идея заключается в построении эмбеддингов и дальнейшем решении задачи многоклассовой классификации. Актуальность задачи обусловлена необходимостью повышения эффективности процессов логистики и учета в сфере закупок, а также сокращения времени и ресурсов, затрачиваемых на классификацию товаров.

Наиболее релевантными работами к данному исследованию оказались [1], [2] [3]. В этих работах исследуется многоклассовая классификация по различным кодам. Наиболее часто встречаются модели — логистическая регрессия и нейронные сети. Также в [1] уделено внимание дообучающимся моделям, таким как GPT-3.5, GPT-4 — они работают лучше остальных с текстами плохого качества, т.е. в которых есть словами или символы неизвестные модели эмбеддингов. В этих статьях при работе с текстом описываются варианты постреления эмбеддингов, фигурируют методы Word2Vec и Glove и различные оболочки, поддерживающие эти методы.

Анализ методов построения эмбеддингов [4] [5] привёл к всё тем же методам Word2Vec[6][7] или GloVe[8]. Эти методы используют нейронные сети и стараются либо предугадать пропущеное слово по контексту, либо восстановить контекст по слову. Стоит выделить предобученные модели, такие как BERT, они адаптированны под специфику языка [9] и дообучаются на входных данных. Таким образом обеспечивают дополнительное преимущество за счёт учёта лингвистических особенностей.

В работе предлагается использовать библиотеку spaCy[10] для построение эмбеддингов. Библиотека основана на вышеописанных методах и имеет предобученные модели, готовые для взаимодействия с

[1] Общероссийский классификатор продукции по видам экономической деятельности, сайт

русским языком. После построения эмбендингов — решается задача классификации, с помощью логистической регрессии. Дальше исследуется качество классификации варьируя глубину классификатора — что и есть основная суть исследования. Для улучшения качества предлагается идея иерархической классификации [11] [12] [13].

2 Постановка задачи

В данной части обсудим формальное теоретическое описание задачи и предлагаемого решения.

Основные термины:

- ОКПД 2 (Общероссийский классификатор продукции по видам экономической деятельности) – система классификации, используемая для каталогизации продукции.
- Эмбеддинги — векторные представления слов. Векторы отражают семантическое значение каждого слова на основе контекста.
- Описание товара — короткий текст, составленные людьми при оформлении продажи товара.
- Префиксы кода длины N (далее "префикс N") — первые N цифр ОКПД2 кода

В данной работе рассматривается задача многоклассовой классификации текстовых описаний товаров для определения их соответствия классам кодам ОКПД2.

2.1 Выборка

Выборка представлена парами "текстовое описание товара — код ОКПД2".

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m},\ \mathbf{x}_i = \ \{\mathtt{t}_i\}_{j=1}^{n} - \texttt{текстовое описание},\ y_i \in \mathbf{Y} = \{1, \dots, k\}. \quad (1)$$

Выборка разбита на обучающую и тестовую части: $\mathfrak{D} = \mathfrak{D}_{train} \bigsqcup \mathfrak{D}_{test}$.

2.2 Ограничения и другие предположения о характере данных

- Количество записей $m \approx 8$ миллионов, количество классов $k \approx 5000$.
- Структура классов несбалансирована: для некоторых классов доступно до 1000 записей, в то время как для других — более 200000.
- Текстовые описания часто содержат узкоспециализированную лексику, жаргонизмы, артикли и числовые значения, что усложняет задачу классификации.

Исходя из особенностей текстовых данных, принимается во внимание, что не все части описаний могут быть одинаково информативными для каждого класса, и необходимо отказаться от некоторых элементов описаний и даже от несущественных классов.

2.3 Определение модели

Используется модель логистической регрессии, она моделирует вероятность принадлежности наблюдения к одному из классов. Обычно она используется для бинарной классификации, но может обобщаться на многоклассовую разными способами — в работе выбрана `OVR` схема (`one-versus-rest`[14]). В случае многоклассовой классификации, модель определяется как:

$$P(y = 1|\mathbf{x}; \theta_k) = \sigma(\mathbf{x}^\top \theta_k), \quad (2)$$

где $\mathbf{x}$ обозначает вектор признаков наблюдения (с предварительно добавленной единицей для учета свободного члена), θ_k — вектор параметров модели для класса k, а $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция. Обучается отдельная логистическая модель для каждого класса, сравнивающая этот класс со всеми остальными классами. Класс с наибольшей предсказанной вероятностью выбирается в качестве итогового предсказания для объекта.

2.4 Функция потерь

Функция потерь для логистической регрессии — это логистическая потеря (log-loss), которая для нашей можели выражается как:

$$\mathcal{L}(\theta_k) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\sigma(\mathbf{x}_i^\top \theta_k)) + (1-y_i)\log(1-\sigma(\mathbf{x}_i^\top \theta_k))\right], \tag{3}$$

где N — число наблюдений в подвыборке, y_i — истинная метка класса для наблюдения i.

2.5 Задача оптимизации

Задача оптимизации для логистической регрессии формулируется как поиск оптимального вектора параметров θ_k, минимизирующего функцию потерь:

$$\theta_k{}^* = \arg\min_{\theta_k} \mathcal{L}(\theta_k). \tag{4}$$

2.6 Критерии качества

Для анализа качества модели используются ROC[15] и Precision-Recall[16] кривые, позволяющие оценить баланс между чувствительностью модели и её способностью корректно классифицировать объекты разных классов. Анализ площади под этими кривыми (`roc-auc` и `pr-auc`) дает количественную оценку эффективности модели.

3 Решение

3.1 Свойства модели или предлагаемого решения

Для анализа текстовых данных и классификации используется модель логистической регрессии, реализованная в библиотеке *scikit-learn*[2]. Логистическая регрессия выбрана за её способность эффективно работать с высокоразмерными данными и выделять линейные зависимости между признаками и целевыми классами. Библиотека поддерживает мультиклассовую классификацию методом `ovr`, а также считает интересующие нас метрики ROC и Precision-Recall.

Для преобразования текстов в векторизованное представление используются эмбеддинги, полученные с помощью библиотеки `spaCy`. Это позволяет преобразовать исходный текст в векторное пространство, где каждое измерение содержит некоторую семантическую информацию о слове, что значительно улучшает качество классификации.

3.2 Описание алгоритма получения решения

Процесс решения задачи классификации состоит из нескольких этапов:

1. Предварительная обработка данных: очистка текста от шума, нормализация и токенизация.
2. Фильтрация данных, для улучшения свойств выборки.
3. Преобразование текстов в векторное представление с использованием эмбеддингов `spaCy`.
4. Обучение модели логистической регрессии на обработанных данных.
5. Оценка качества модели с roc-auc и pr-auc.

4 Вычислительный эксперимент

4.1 Цель эксперимента

Цель эксперимента построить модель логистической регресии и исследовать её работу в зависимости от глубины классификатора. Также необходимо провести два эксперимента с применением иерархической классификации: мы достаточно хорошо должны уметь прогнозировать первые несколько цифр кода, а

[2]По этой ссылке находится используемая модель

так как данные имеют иерархическую структуру, хотелось бы использовать результаты предсказания первых цифр при предсказании дальнейших.

- Эксперимент 1: Будем добавлять предсказанные цифры в вектор признаков. Так как эмбеддинги — это вектора, каждый элемент которых - число от 0 до 1, а номер класса - число от 1 до 1000 или даже больше, модель будет обращать внимание на этот признак и учитывать его.
- Эксперимент 2: Будем разделять выборку на подклассы, по результатам классификации по первым цифрам. Далее для каждого класса - обучаем свою модель. Эта идея похожа на предыдущую, она по сути представляет собой более строгое использование результата предыдущей классификации.

4.2 Описание постановки и условий эксперимента

- В качестве реализации модели лог регрессии выбрана библиотека `sklearn` и её `sklearn.linear_model.LogisticRegression`
- Для построения эмбеддингов выбрана библиотека `spaCy` и её предобученная модель `ru_core_news_lg`. Стоит оговориться:
  - Проводились отдельные эксперименты с библиотекой `gensim`, для получения другой вариации эмбеддингов. Но результаты классификации оказались значительно хуже[3].
  - Размерность векторов эмбеддингов сильно влияет на классификацию. В экспериментах выбрана размерность 300, как оптимальная для использования `large` модели из `spaCy`.
  - Выбор именно этой предобученной модели, был обоснован сравнением качества эмбеддингов полученных с помощью `large` и `small` моделей.

4.3 Описание данных и их обработка

Изначально данные — это пары значений: текстовое описание товара и его ОКПД2 код. Эти описания были составлены людьми, содержат орфографические ошибки, лишние символы, артикулы, цифры и много другое. Необходимо произвести предобработку:

- Данные привели к однотипному формату: убрали склонения, заглавные буквы, почистили от незначащих символов, артикулов.
- Данных очень много, так что из-за вычислительной сложности эксперимента было решено выделить несколько классов для анализа их классификации. [4]
- Данные прошли фильтрацию для избавления от слишком мелких классов, чтобы снизить дизбаланс классов и улучшить качество классификации.

Итоговая подвыборка после предобработки, используемая в большинстве экспериментов, сформирована и загружена отдельно.

[3] Здесь представлены все результаты экспериментов

[4] Выбор пал на 1,17,33,45,58,81,86 классы по первым двум цифрам кода. Это : 01 — Продукция и услуги сельского хозяйства и охоты, 17 — Бумага и изделия из бумаги, 33 — Услуги по ремонту и монтажу машин и оборудования, 45 — Услуги по оптовой и розничной торговле и услуги по ремонту автотранспортных средств и мотоциклов, 58 — Услуги издательские, 81 — Услуги по обслуживанию зданий и территорий, 86 — Услуги в области здравоохранения

4.4 Ход эксперимента

0. Для всех данных построены эмбеддинги.
1. Эксперимент 1, выборка 1[5].
   (a) По обучающей выборке обучены модели логистической регрессии, в качестве ответов выбраны префиксы кодов длиной от 3 до 7. (соответственно обучено 5 моделей)
   (b) Для каждой модели и каждого класса построены presign-recall и roc - кривые
2. Эксперимент 2, выборка 2[6].
   (a) По обучающей выборке обучены модели логистической регрессии, в качестве ответов выбраны префиксы кодов длиной от 3 до 7. (соответственно обучено 5 моделей)
   (b) Для каждой модели и каждого класса построены presign-recall и roc - кривые
3. Для выборки 2 проведено два эксперимента с проверкой иерархической идеи улучшения:
   (a) Полученные результаты после классификации по префиксу кода длиной 3 добавлены в вектор признаков. Теперь по увеличенному вектору признаков (эмбеддинги + predicted-1-step) обучаются модели логистической регрессии,, в качестве ответов выбраны префиксы кодов длиной от 3 до 6.
   (b) Все данные разделены по подклассам, определённым после классификации по префиксу кода длиной 3. Для каждого подкласса обучаются отдельные модели лог регрессии, в качестве ответов выбраны префиксы кодов длиной от 3 до 6.
4. Для анализа ошибок этих экспериментов - подсчитаны количества неправильных ответов моделей, в зависимости от длины префикса кода.

4.5 Анализ полученных результатов

- По результатам пункта 2 раздела 4.4 была замечена основная тенденция обучения: вне зависимости от глубины, большие классы хорошо отделяются от остальных, точность падает на более длинных префиксах кода, но до допустимых значений ≈ 0.5. Для маленьких классов, модель часто не справляется.
- На рисунках 1 и 2 - PR и ROC кривые для всех классов префикса кода длиной 3. Видим, что все ROC-кривые показывают отличные результаты. Видим, что все PR-кривые показывают хороший результат, кроме 3 и 4 классов, соответствующих "услугам по ремонту и обслуживанию машин и спец техники". После анализа, выявлено содержание множества некачественных текстов, среди описаний товаров этих классов, по которым сроятся некачественные эмбеддинги. В дальнейших экспериментах был убран этот класс из рассмотрения, до возможности дообучения эмбеддингов на наших данных.

- По результам пункта 5, наблюдаем подтверждение тенденции качественного отделения больших классов и видим улучшение результатов классификации при исключении классов с плохими данными. На рисунке 3 - показаны pr-кривые для всех классов префикса кода длиной 3 выборки 2, видим хорошие показатели разделения.
- На рисунке 4.5 - показаны pr-кривые для одного из классов в зависимости от глубины классификации. Видим закономерное ухудшение результата классификации при увеличении длины префикса кода, при этом даже при предсказании целого кода результаты удовлетворительные.

- По результатам экспериментов с проверкой иерархической идеи улучшения классификации получена сводная таблица 1. В ней показаны количества ошибок (в тысячах), на выборке 2 в зависимости от метода: стандартный, с разбиением выборки на подклассы, с добавлением нового признака. Размер выборки 2 — 736 тыс. записей.
- В таблице 1 наблюдается рост количества ошибок с увеличением длины префикса. Также видим, что предлагаемые методы иерархической классификации улучшают результаты — наблюдаем уменьшение количества ошибок примерно на 5%. Чуть лучше себя показывает метод с разбиением на подклассы, как и предполагалось.

[5]Подвыборка включающая 33,45,81,86 классы по первым двум цифрам
[6]Подвыборка включающая 1,17,33 классы по первым двум цифрам

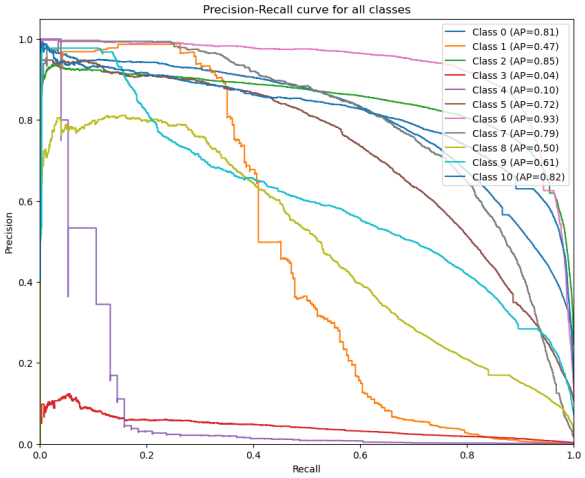


Рис. 1: PR-кривые для всех классов, префикс 3, выборка 1

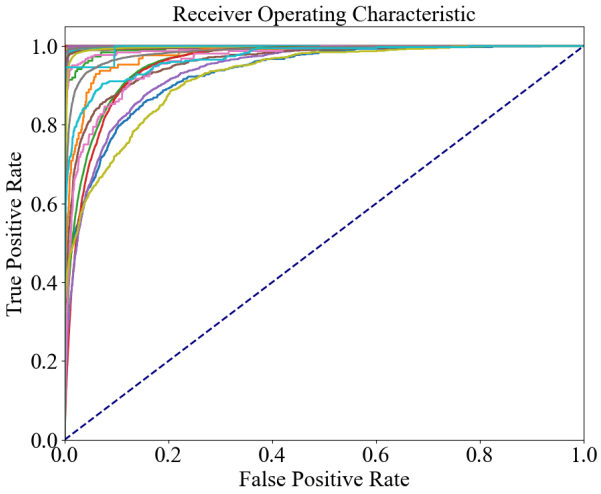


Рис. 2: ROC-кривые для всех классов, префикс 3, выборка 1

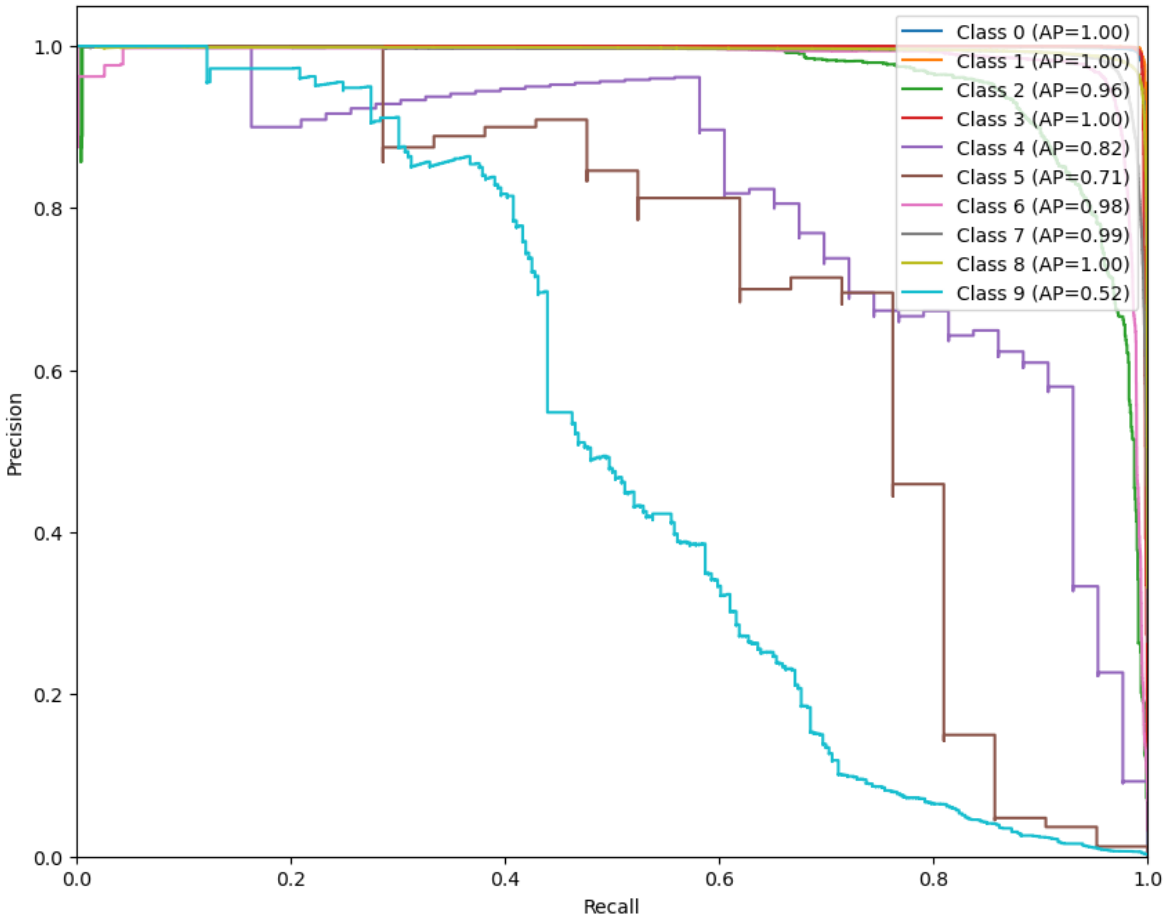
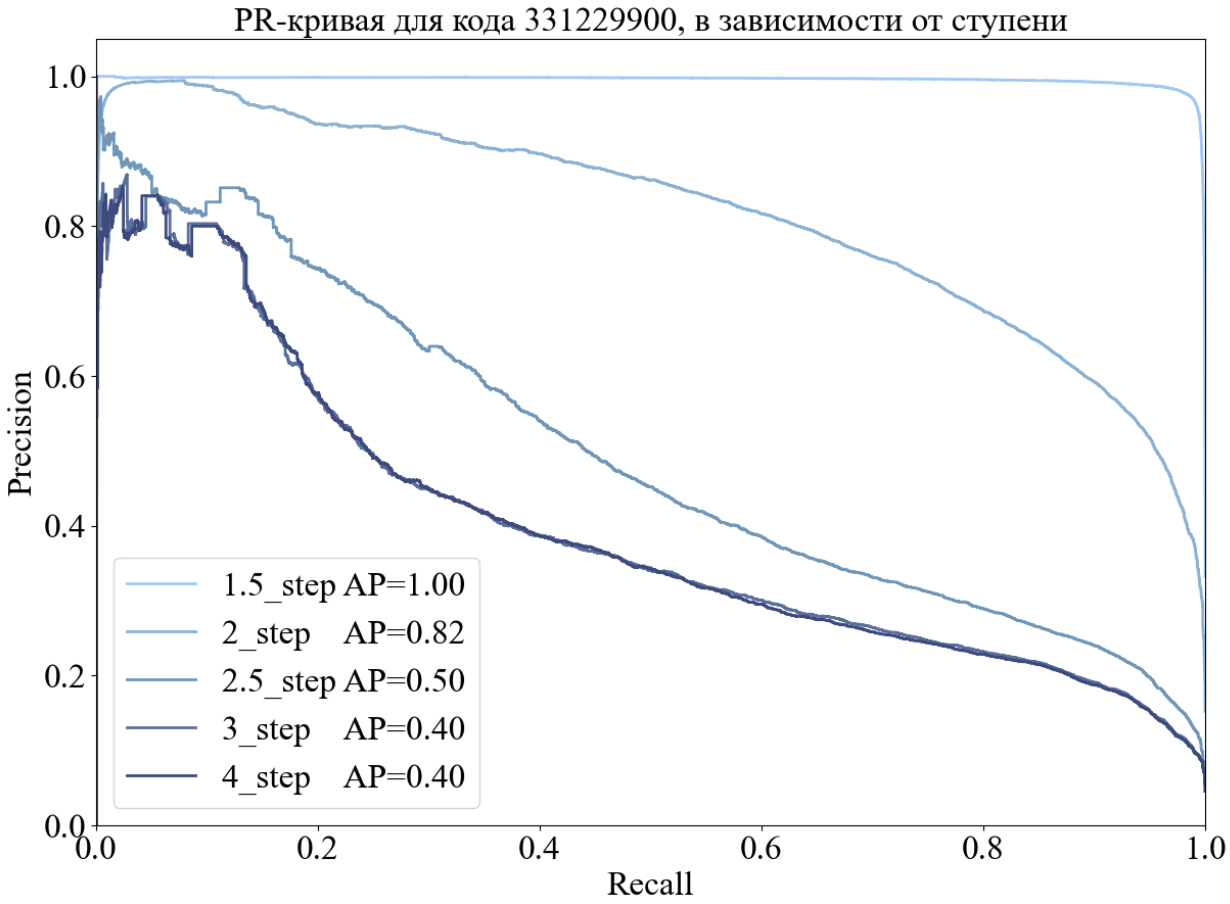


Рис. 3: PR-кривые для всех классов, префикс 3, выборка 2



Метод	Стандартный	С разбиением на подклассы	С добавлением признака
Префикс 4	90	**86**	87
Префикс 5	106	**101**	103
Префикс 6	120	**112**	113

Таблица 1: Количество ошибок на выборке 2 в зависимости от метода и длины префикса

4.6 Выводы

- Модели были построены на различных выборках, и результаты показали, что для крупных классов модель демонстрирует высокую точность, что соответствует целям исследования.
- Анализ зависимости качества классификации от глубины классификатора подтвердил ожидаемые тенденции: с увеличением глубины точность падает, но остаётся в приемлемых пределах.
- Реализация и тестирование улучшенных методов классификации показали снижение количества ошибок на 5%, что указывает на положительное влияние предложенных усовершенствований.

Все эксперименты поддерживают повторяемость и могут быть проверены и проведены любым желающим. Необходимые скрипты и дополнительные материалы доступны на GitHub проекта [17].

5 Заключение

В рамках данной работы был предложен и реализован алгоритм для классификации товаров по кодам ОКПД2. Разработанная модель основывается на использовании текстовых эмбеддингов и логистической регрессии, что позволило эффективно решать задачу многоклассовой классификации. Было проведено тщательное исследование модели, в ходе которого анализировалось влияние глубины классификатора на

качество результатов, а также изучено, как различные гиперпараметры влияют на производительность модели.

Кроме того, в работе были исследованы различные способы улучшения модели, включая модификации алгоритма классификации и применение иерархической структуры классификатора, что позволило улучшить точность классификации, особенно на более глубоких уровнях.

Возможные пути для дальнейших улучшений модели включают:

- Улучшение качества текстовых эмбеддингов, возможно, за счет использования более совершенных моделей нейронных сетей или дополнительного дообучения на специфических данных.
- Интеграция и усовершенствование предложенных методов улучшения классификации для создания более робустной системы.
- Работа над несбалансированностью классов, что может включать техники ресемплинга или специализированные функции потерь.

Список литературы

[1] Ignacio Marra de Artiñano, Franco Riottini Depetris, and Christian Volpe Martincus. Automatic product classification in international trade: Machine learning and large language models. Jul 2021. Available at http://dx.doi.org/10.18235/0005012.

[2] David D. Lewis et al. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5, 2004.

[3] Hele-Mai Haav. Assessment of hs code correctness, 2021.

[4] Ying Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Word embeddings: A survey. *arXiv preprint arXiv:1807.04606*, 2018.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc. 2009.

[6] Kenneth Ward Church. Why are some papers cited more than others? *Natural Language Engineering*, 23(1):155–162, 2017.

[7] Giovanni Di Gennaro, Amedeo Buonanno, and Francesco A. N. Palmieri. Considerations about learning word2vec. *The Journal of Supercomputing*, 77(11):12320–12335, 2021.

[8] Eduardo Muñoz. Introduction to natural language processing: Word embeddings sentiment analysis with python, 2020.

[9] Eunchan Lee, Changhyeon Lee, and Sangtae Ahn. Comparative study of multiclass text classification in research proposals using pretrained language models. *Applied Sciences*, 12(9):4522, 2022.

[10] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[11] Wenhan Liang Rundong Liu. Recent advances in hierarchical multi-label text classification: A survey, 2022.

[12] Yannick Versley. Hierarchical classification of text documents. In *Notes of the Lecture at the European Summer School on Logic, Language and Information (ESSLLI)*, volume 2006. Citeseer, 2006.

[13] Fabrizio Sebastiani. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 34(1):1–50, 2020.

[14] Carmen Jiménez-Mesa, Ignacio Alvarez Illán, Alberto Martín-Martín, Diego Castillo-Barnes, Francisco Jesus Martinez-Murcia, Javier Ramírez, and Juan M. Górriz. Optimized one vs one approach in multiclass classification for early alzheimer's disease and mild cognitive impairment diagnosis. *IEEE Access*, 8:96981–96993, 2020.

[15] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[16] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. General evaluation measures for document organization tasks. *Springer*, 2018, 2018.

[17] Sergey Firsov. Classification of products according to okpd 2 codes. project for the course m1p. https://github.com/intsystems/2024-Project-142, 2024.