Классификация товаров по ОКПД2 кодам

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

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

Вознюк Анастасия Евгеньевна
Антиплагиат
МФТИ
vozniuk.ae@phystech.edu

Аннотация

Исследование направлено на решение задачи классификации товаров по кодам Общероссийского классификатора продукции по видам экономической деятельности (ОКПД 2) с использованием кратких текстовых описаний. Коды представляют собой детализированную систему категоризации продуктов и услуг по видам экономической деятельности. Основная цель — повышение точности и сокращение ресурсозатратности классификации, за счёт анализа влияния глубины ОКПД 2. Для достижения этих целей предлагается метод построения текстовых эмбеддингов. Задача усложняется необходимостью предварительной обработки данных для перевода исходных описаний в стандартизированные короткие тексты, адаптированные для анализа. Используются данные государственных закупок по ФЗ[0]44 за 2022 год. Новизна работы заключается в применении методов машинного обучения к индустриальной задаче, что обещает улучшение в процессах логистики, учёте и анализе в сфере закупок.

Ключевые слова OKPD 2 code · text analysis · multiclass classification · embeddings · logarithmic regression

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. *(тут нужна прямая ссылка где требуется, но у меня индустриальная задача, которая буквально дана как нужная в фирме Всеволода Михайловича)* [1]

Основные термины:

- ОКПД 2 (Общероссийский классификатор продукции по видам экономической деятельности) – система классификации, используемая для каталогизации продукции.
- Эмбеддинги — векторные представления слов. Вектора[2]отражают семантическое значение каждого слова на основе контекста.
- Описание товара — короткий текст, составленные людьми при оформлении продажи товара.

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

- Ступени классификатора — негласное разделение кода на части, удобное при анализе качества от глубины. К примеру — код 12.34.56.789, 1-я ступень — 12, 2-я ступень — 1234, и так далее.

При обзоре литературы на тему многоклассовой классификации наиболее релевантны оказались работы Automatic Product Classification in International Trade[5], RCV1: A New Benchmark Collection for Text Categorization Research[4] и Assessment of HS Code Correctness, Haav, Hele-Mai[3]. В этих работах исследуется многоклассовая классификация по различным кодам. Наиболее популярные модели для решения — логарифмическая регрессия и нейронные сети, также в [5] уделено внимание дообучающимся моделям, таким как GPT-3.5, GPT-4 — они работают лучше остальных с текстами плохого качества, со словами неизвестными моделям эмбеддингов. В этих статьях при работе с текстом описываются варианты построения эмбеддингов, фигурируют методы Word2Vec и Glove и различные оболочки, поддерживающие эти методы.

Анализ методов построения эмбеддингов привёт к всё тем же методам Word2Vec или GloVe. Эти методы используют нейронные сети и стараются либо предугадать пропущеное слово по контексту, либо восстановить контекст по слову. Опираемся на статьи: Word2vec[1], Introduction to Natural Language Processing[6], Considerations about learning Word2Vec [2].

В работе предлагается использовать библиотеку spaCy[2] для построение эмбеддингов. Библиотека основана на вышеописанных методах и имеет предобученные модели, готовые для взаимодействия с русским языком. После построения эмбендингов — решаем задачу классификации, с помощью логарифмической регрессии. Дальше исследуем качество классификации варьируя глубину классификатора — что и есть основная суть исследования. Для улучшения качества предлагается идея иерархической классификации: разделив код на ступени по несколько цифр, предсказываем их по очереди, при предсказании более глубоких ступеней — используем результаты предыдущих.

2 Постановка задачи

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".

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m},\ \mathbf{x}_i = \{token\}_{j=1}^{n} - array\ of\ words,\ y_i \in \mathbf{Y} = \{1, \dots, k\}. \tag{1}$$

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 признаки 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`). В случае многоклассовой классификации, модель определяется как:

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

[2] Официальная документация spaCy

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:

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

11

2.6 Критерии качества 12

Для анализа качества модели используются ROC и Precision-Recall 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`. 13 Логистическая регрессия выбрана за её способность эффективно работать с высокоразмерными данными и выделять линейные зависимости между признаками и целевыми классами, а также на основе анализа литературы. Библиотека поддерживает мультиклассовую классификацию методом `ovr`, а также интерпретирует результаты в интересующих нас форматах (ROC и Precision-Recall кривые).

Для преобразования текстов в векторизованное представление используются эмбеддинги, полученные с помощью библиотеки `spaCy`. Это позволяет преобразовать исходный текст в плотное 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.

3.3 Свойства алгоритма 14

Алгоритм логистической регрессии обладает следующими свойствами:

- Высокая интерпретируемость результатов благодаря линейной природе модели.
- Способность эффективно обрабатывать разреженные данные, что часто встречается в текстовых задачах.
- Возможность оценки вероятности принадлежности к классам, что позволяет анализировать не только итоговую классификацию, но и уверенность модели в своих предсказаниях.

4 Вычислительный эксперимент

4.1 Цель эксперимента

15 Построить эмбеддинги по текстовым описаниям. Построить модель логарифмической регресии, для классификации по эмбеддингам. Исследовать работу модели в зависимости от глубины классификатора. Проверить идеи применения иерархической классификации: мы достаточно хорошо должны уметь прогнозировать первые ступени классификатора, а так как данные имеют иерархическую структуру, хотелось бы использовать результаты предсказания первых ступеней при предсказании дальнейших.

16 17

- Идея 1: Будем добавлять предсказанные ступени 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.
- Идея 2: Будем разделять выборку на подклассы, по результатам классификации по первым ступеням. 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].

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[5].
- В данных добавлены промежуточные ступени классификатора: полный код — это 9 цифр, пример 12.34.56.789. Разбиваем его на числовые значения по ступеням: 1 ступень — 12, ступень 1.5 — 123, ступень 2 — 1234, ступень 2.5 — 12345 и т.д..

[3] Здесь представлен эксперимент с построением эмбеддингов через `gensim` и его результаты
[4] Файл с результатами lg модели (прямое сравнение добавлю позже, будет с bert ещё)
[5] Файл с предобработкой

- Данных очень много, так что из-за вычислительной сложности эксперимента было принято выделить несколько классов для анализа их классификации. Выбор пал на 1,17,33,45,58,81,86 классы по первой ступени.
- Данные прошли фильтрацию для избавления от слишком мелких классов, чтобы избегать дизбаланса классов и улучшать качество классификации.

Итоговая подвыборка после предобработки, используемая в большинстве экспериментов, сформирована и загружена отдельно[6].

4.4 Ход эксперимента

Я не знаю почему тут такие пробелы между строками

1. Для всех данных построены эмбеддинги и добавлены отдельной колонкой в датасет. 18
2. Взята подвыборка включающая 33,45,81,86 классы по 1-ой ступени (далее - выборка 1). Для неё произведено разбиение на train/test части.
3. По обучающей выборке обучены модели логарифмической регрессии, в качестве меток выбраны коды ступеней 1.5, 2, 2.5, 3. (соответственно обучено 4 модели)
4. Для каждой модели и каждого класса построены presign-recall и roc - кривые
5. Взята подвыборка включающая 1,17,33 классы классы по 1-ой ступени (далее - выборка 2). Для неё проведены пункты 3-4.
6. Для этой же подвыборки проведено два эксперимента с проверкой иерархической идеи улучшения:
   (a) Полученные результаты после классификации по ступени 1.5 добавлены в вектор признаков. Теперь по увеличенному вектору признаков (эмбеддинги + predicted-1-step) обучаются модели логарифмической регрессии, в качестве меток выбраны коды ступеней 2, 2.5, 3.
   (b) Все данные разделены по подклассам, определённым после классификации по ступени 1.5. Для каждого подкласса обучаются отдельные модели лог регрессии, в качестве меток выбраны коды ступеней 2, 2.5, 3.
7. Для анализа ошибок этих экспериментов - подсчитаны количества неправильных ответов моделей, в зависимости от ступеней.

4.5 Анализ полученных результатов

- По результатам пункта 2[7] была замечена основная тенденция обучения: вне зависимости от ступени, большие классы хорошо отделяются от остальных, точность падает на более глубоких ступенях, но до допустимых значений. Для маленьких классов, модель часто не справляется, что оказалось и не необходимо, после уточнения постановки задачи.
- На рисунках 1 и 2 - PR и ROC кривые для всех классов ступени 1.5. Видим, что все ROC-кривые показывают отличные результаты, в дальнейшем их не будем приводить, они всегда показывают хорошие результаты из-за дизбаланса классов и отличных результатов классификации по большим классам. 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[8], 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 ступени 1.5 выборки 2, видим хорошие показатели разделения.

[6] Для воспроизведения экспериментов рекомендуется использовать этот файл
[7] Эксперимент и результаты можно посмотреть тут
[8] Эксперимент и результаты можно посмотреть тут

19

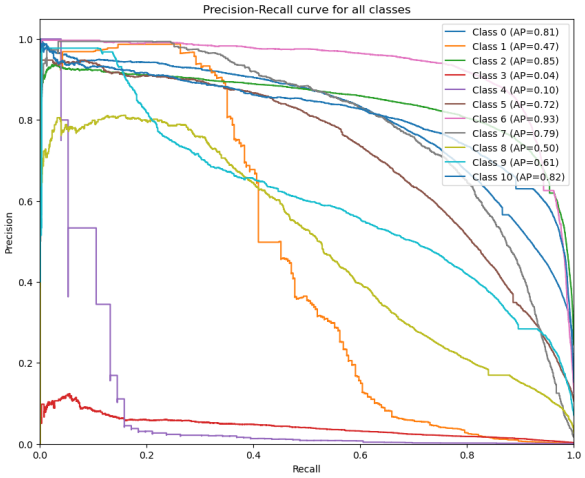


Рис. 1: PR-кривые для всех классов, ступень 1.5, выборка 1

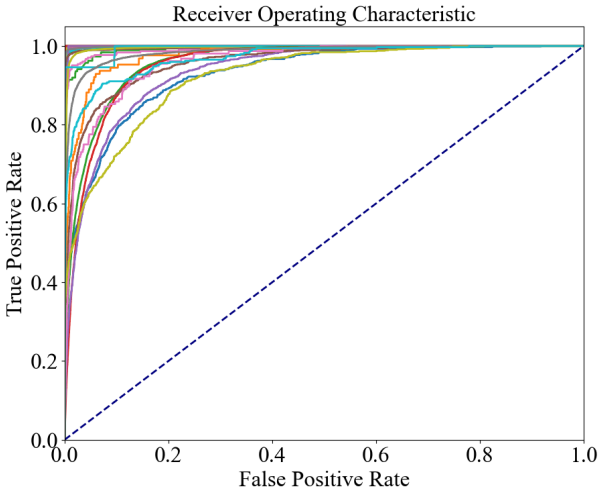


Рис. 2: ROC-кривые для всех классов, ступень 1.5, выборка 1

- На рисунке 4.5 - показаны pr-кривые для одного из классов в зависимости от глубины классификации. Видим закономерное ухудшение результата классификации при увеличении ступени, при этом даже на последней ступени результаты удовлетворительные.

- По результатам экспериментов [9] 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.

Метод	Стандартный	С разбиением на подклассы	С добавлением признака
Ступень 2	90	**86**	87
Ступень 2.5	106	**101**	103
Ступень 3	120	**112**	113

Таблица 1: Количество ошибок на выборке 2 в зависимости от метода и ступени 20

- В таблице 1 наблюдается предполагаемый рост количества ошибок с увеличением номера ступени. 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.

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 ошибок на 5%, что подтверждает эффективность предложенных усовершенствований.

21

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

[9] Эксперименты и результаты можно посмотреть тут и тут

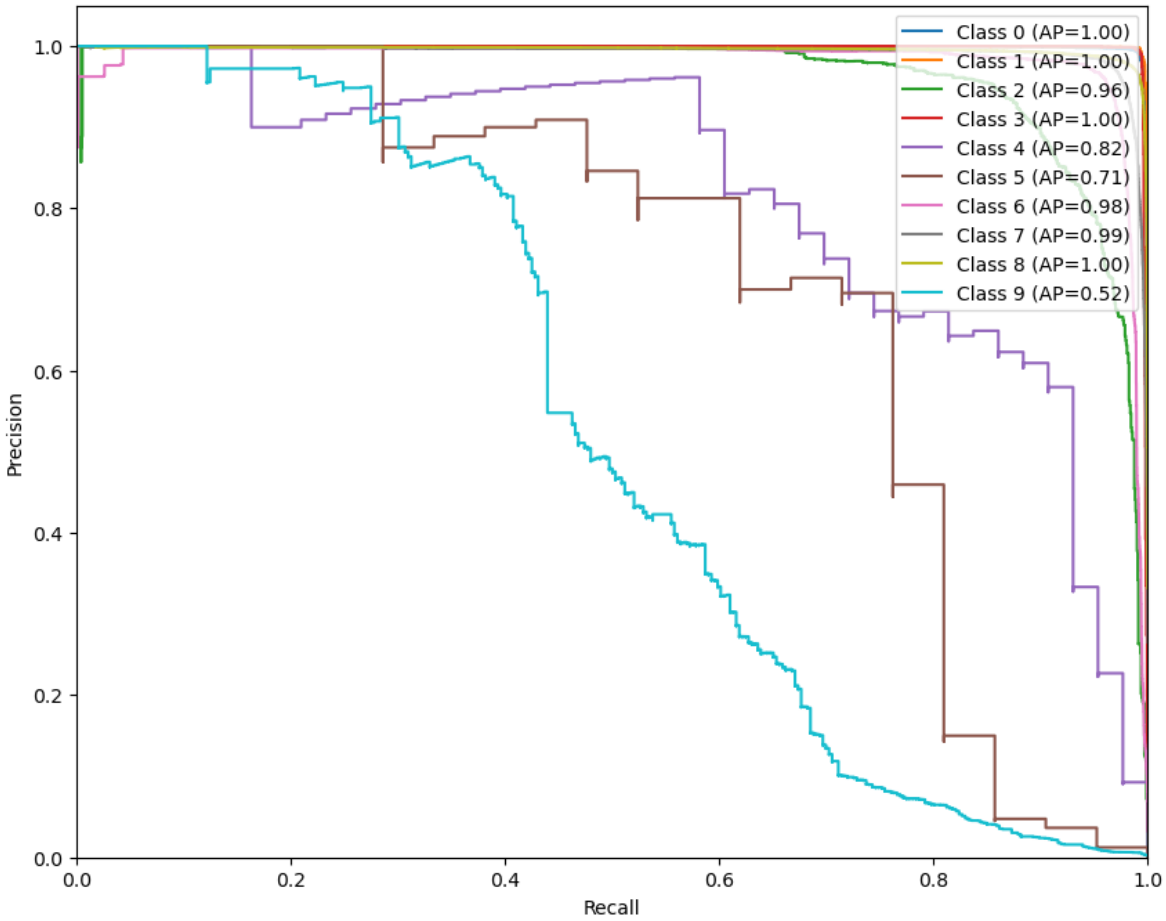


Рис. 3: PR-кривые для всех классов, ступень 1.5, выборка 2

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.

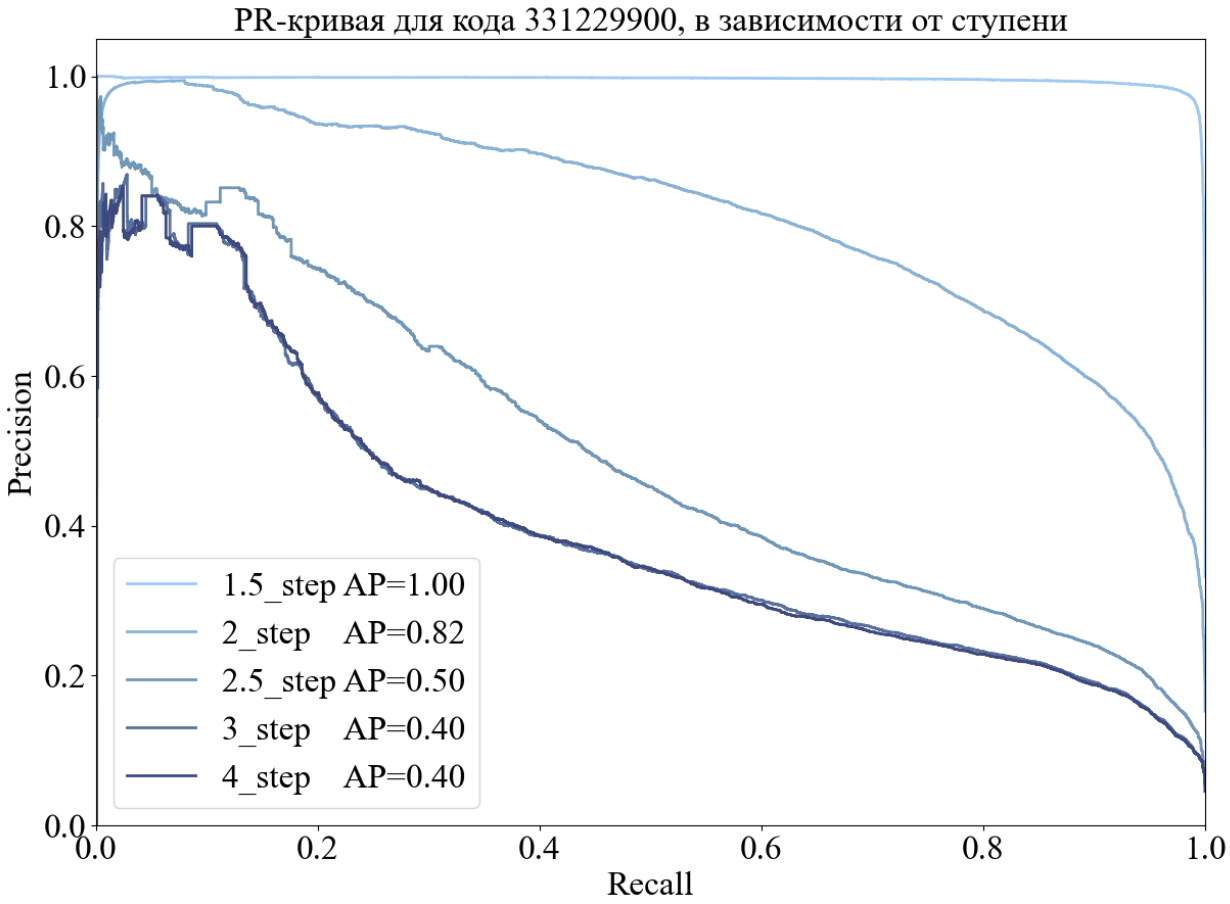
Возможные пути для дальнейших улучшений модели включают:

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

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



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

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

[5] Ignacio Marra de Artiñano, Franco Riottini Depetris, and Christian Volpe Martincus. Automatic product classification in international trade: Machine learning and large language models. *Not provided*, Not provided(Not provided):Not provided, Jul 2021. Available at http://dx.doi.org/10.18235/0005012.

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

0. Федеральному закноу, может лучше?

1. Вбейте в scholar что-нибудь типа "product classification code intelligence"
 и найдите относительно свежую релевантную обзорную статью, где говорят что это важно

2. Правильно "векторы", вектора - жаргонизм, в статье лучше не писать

3. Обычно названия работ не указывают, а просто ставят ссылки типа:
" наиболее релевантные работы~\cite{blablabla1,blablabla2}.
+: лучше не писать "оказались работы". Вы здесь презентуете результат вашего исследования,
а не процесс его получения. Поэтому лучше что-то типа
"Наиболее релевантными работами к данному исследованию являются~\cite{...."

4. Наиболее популярные - оценочное суждение. Либо нужна ссылку на мета-анализ, где проверяют, что
метод является действительно популярным, либо заменить это предложение на более нейтральное

5. привёт? В любом случае, нужно переписать более формально, с оглядкой на комментарий 3.

6. Методы, рассматриваемые в данной работе, опираются на статьи

7. Вбейте в гугл "how to cite spacy", первая сслка даст цитату в нормальном виде, не через footnote

8. Кавычки неправильные, посмотрите презентацию с типовыми ошибками, которую я кидал в чат в марте

9. Имеет смысл дать здесь цитату на какую-то классическую работу по ML, где есть этот термин

10. \theta должна быть жирной, раз это вектор.
 В латехе символы из этого алфавита монжо сделать жирными так
\boldsymbol{theta}

11. Указать по какому множеству идет оптимизация

12. Комментарий спорный, но я бы не делал таких мелких подсекций. Если хочется поделить текст,
можно попробовать делать \paragraph вместо \subsection

13. Нужна ссылка на пакет

14. Не уверен что нужно целая секция, которая перечисляет
свойства классической ML-модели. Если какие-то свойства хочется подчеркнуть, наапишите об этом абзац

15. а. Так не пишут, заголовок секции не является началом предложения основного текста
    б. Тут немного намешано. Цель эксперимента у вас --- проанализировать работоспособность
     предложенного метода (или что-то в этом роде)

16. Непонятно что такое "идея" здесь. Если было проведено два различнхы эксперимента,
то так и напишите

17. эбмддинги --- это (нужно длинное тире)

18. Выборка

19. Для технического отчета ссылки на промежуточные результаты - это хорошо.
Для статьи - лучше сделать одну ссылку на репозиторий и все ссылки на файлы и результаты указать
в readme репозитория

20. Здесь нужно учесть std при проведение нескольких экспериментов

21. Видно, что проделана большая работа. Пока эксперииимент выглядит немного сыро - очень
 много ненужных деталей,
в которых можно запутаться. Часть вещей лучше написать более формальным отчетным языком.
 Посмотрите на работы коллег и прошлогодние работы.