Классификация товаров по ОКПД2 кодам

A Preprint

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

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

Abstract

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 с использованием нейросетевых технологий. Задача усложняется необходимостью предварительной обработки данных для перевода исходных описаний в стандартизированные короткие тексты, адаптированные для анализа. Используются данные государственных закупок по ФЗ 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.

Keywords OKPD 2 code · text analysis · task of classification

1 Введение

Целью данного исследования является разработка и апробация метода классификации товаров по кодам ОКПД 2 (ОКР [2024]), используя краткие текстовые описания. Актуальность задачи обусловлена необходимостью повышения эффективности процессов логистики и учета в сфере закупок, а также сокращения времени и ресурсов, затрачиваемых на классификацию товаров.

Объектом исследования выступают любые товары, для которых возможна классификация по ОКПД 2 кодам (детализированной системе категоризации продукции и услуг по видам экономической деятельности). Проблема заключается в разработке метода, позволяющего автоматизировать этот процесс с высокой точностью и полнотой классификации, устойчиво относительно формата входных данных, и в исследовании характеристик этого метода (по указанным параметрам) от глубины классификации.

Задача классификации по кодам разобрана здесь (добавить). В дальнейшей работе также будем опираться на курс лекций Воронцова К.В. и книги Гудфеллоу et al. [2016] и Montani [2019]. Текстовый эмбеддинг — веторное представление слова. Вектора отражают семантическое значение каждого слова на основе контекста. Наиболее часто они получаются при помощи методов Word2Vec или GloVe. Эти методы используют нейронные сети и стараются либо предугадать пропущеное слово по контексту, либо восстановить контекст по слову. Опираемся на эти статьи: Alammar [2019] и Muñoz [2020]. Мы пользуемся библиотекой spaCy spa, основанной на вышеописанных методах и имеющей предобученные модели, готовые для взаимодействия с русским языком и короткими текстами. После построения эмбендингов — решаем задачу классификации, сопоставляем вектора и классы ОКПД 2. Дальше исследуем качество классификации варьируя глубину классификатора — что и есть основная суть исследования.

Используются данные государственных закупок по ФЗ 44 за 2022 год. Это позволяет оценить работу алгоритма в условиях большого объема и разнообразия данных. Исследование включает в себя подготовку

данных, построение модели классификации и ее тестирование с целью определения оптимальных параметров для достижения максимальной точности классификации в зависимости от необходимой глубины класификации. Рабочий процесс описывает последовательные шаги от предварительной обработки текстов до оценки результатов классификации.

В заключение, данное исследование представляет собой вклад в развитие методов машинного обучения и их применение к решению практических задач классификации товаров, что имеет важное значение для сферы государственных закупок и управления цепочками поставок.

2 Постановка проблемы

Дана выборка для задачи многоклассовой классификации, с количеством классов K:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m},\ \mathbf{x}_i = text\ description,\ y_i \in \mathbb{Y} = \{1, \ldots, K\},$$

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

Будет использоваться линейная модель классификации, $A = \{\ g(x,\theta) \mid \theta \in \mathcal{R}\ \}$, где $g(x,\theta) = sign \sum_{j=1}^{n} \theta_j f_j(x)$.

Используется квадратичная функция потерь: $\mathcal{L}(a,x) = (a(x) - y(x))^2$, где y — значения контрольной выборки.

Откуда получаем функцию Эмпирического риска : $Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(a, x_i)$

И будем решать задачу оптимизации — минимизации эмпирического риска

$$\mu(X^l) = arg \min_{a \in A} Q(a, X^l)$$

Критерий качества в задаче нашей классификации - попадание конкретного товар в свою категорию и в правильный ОКПД 2 код.

3 Планирование вычислительного эксперимента

Цель эксперимента — исследование алгоритма классификации, построенного с помощью текстовых эмбендингов, в зависимости от глубины классификации.

Выборка данных из госзакупок, представляет собой набор объектов — текстовых описаний товаров, и их признаков — код их ОКПД2 классификации. Описания некоторых товаров слишком специфичны, для чего их укоротили и стандартизировали.

Пример итогового описания и кода: Яйца куриные в скорлупе свежие 01.47.21.000

Выборка разделена на обучающую и тестовую часть, размерами 6 и 2 миллиона записей соответственно.

Используя sklearn и spaCy строятся сначала текстовые эмбендинги для слов, далее по ним решается задача классификации. Варьируется необходимая глубина классификации и исследуется в зависимости от этого количество ошибок при классификации.

Таблицы - точность алгоритма в зависимости от ступени класификатора: в процентном соотношении для 1,2,3,4 ступени. Груфик по этой таблице.

4 Базовый алгоритм

Для базового алгоритма выбрана лог регресия. В качестве набора данных взята часть выборки — первый миллион записей. По ним строятся эмбендинги и с помощью sklearn решается задача лог регрессии. По полученным результатам строим ROC кривую.

Код и сама выборка находятся в соответствующих файлах проекта.

Здесь ROC кривая для первого класса после классификации:

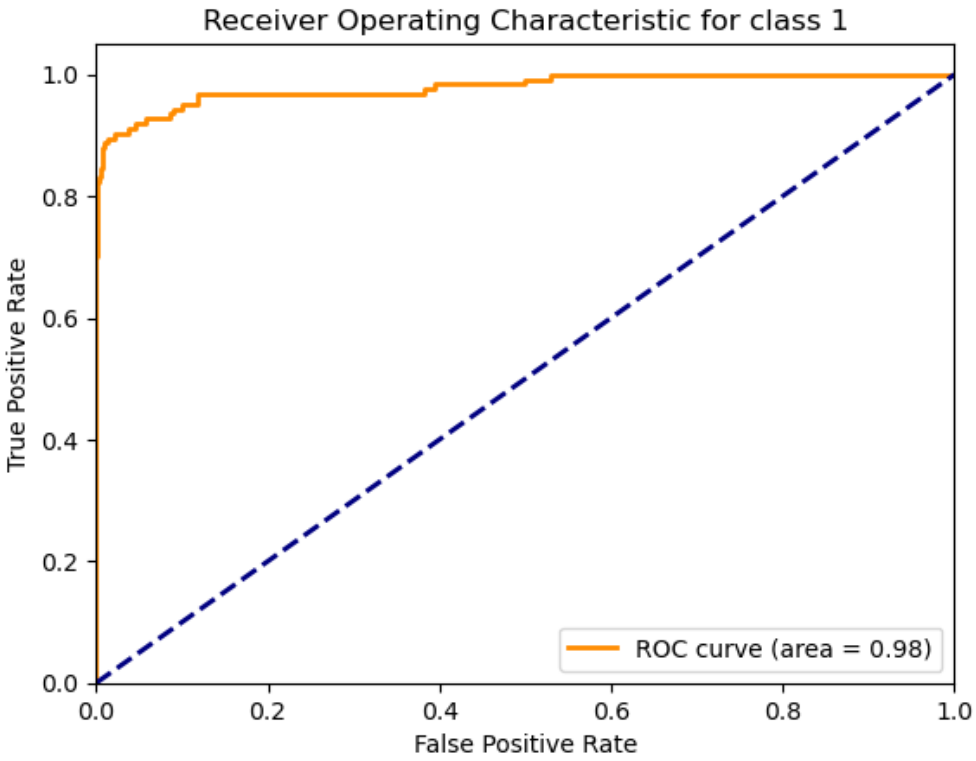

Рис. 1: ROC кривая для одного из классов после базового эксперимента

Список литературы

Общероссийский классификатор продукции по видам экономической деятельности (ОКПД). `https://classifikators.ru/okpd`, 2024. Доступно онлайн.

Ян Гудфеллоу, Йошуа Бенджио, and Аарон Курвилль. Глубокое обучение. MIT Press, 2016.

Ines Montani. Advanced NLP with spaCy: A Practical Guide to Advanced Natural Language Processing. Independent, 2019.

Jay Alammar. The illustrated word2vec, 2019. URL `https://jalammar.github.io/illustrated-word2vec/`.

Eduardo Muñoz. Introduction to natural language processing: Word embeddings sentiment analysis with python, 2020. URL `https://edumunozsala.github.io/BlogEms/jupyter/nlp/classification/embeddings/python/2020/08/15/Intro_NLP_WordEmbeddings_Classification.html`.

Официальная документация spacy. URL `https://spacy.io/`.