Федеральное государственное автономное образовательное
учреждение высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**
Факультет компьютерных наук

КУРСОВАЯ РАБОТА

Разработка алгоритма расчета индекса этичности компании на основе текстовых открытых данных

по направлению подготовки Прикладная математика и информатика
образовательная программа «Финансовые технологии и анализ данных»

Работу выполнил:
Студент группы мФТиАД22
Ященко Михаил Андреевич

_____

Научный руководитель:
Кандидат экономических наук
Сторчевой Максим Анатольевич

_____

Москва 2023

Аннотация

Сейчас, когда говорят о оценке этичности компаний, чаще всего речь идет про ESG. Это важно, для снижения рисков, как самих компаний, так и их клиентов, улучшения репутации компании: если компания придерживается принципов ESG, это может привлечь привлечь больше инвесторов и партнеров. Также ESG оценивает, как компании влияют на качество жизни людей и общества в целом, и окржующую среду. На данный момент есть несколько рейтитингов ESG. Но с этими рейтингами есть ряд проблем: рейтинги не всегда согласуются между собой из-за разных методик оценки, чаще всего это трудоемкая ручная работа, требующая много часов, в текущих рейтингах не так много компаний. Поэтому встал вопрос об автоматизации расчета индекса этичности компаний. Для этого мы решили использовать нефинансовые отчеты компаний с сайта РСПП и на их основе разработать собственный алгоритм расчета показателей ESG.

Now, when we talk about assessing the ethics of companies, most often we are talking about ESG. This is important to reduce the risks of both the companies themselves and their customers, to improve the reputation of the company: if the company adheres to the principles of ESG, it can attract more investors and partners. ESG also assesses how companies affect the quality of peoples life and society, and the environment. There are several ESG ratings at the moment. But there are a number of problems with these ratings: they do not always agree with each other due to different evaluation methods, most often it is laborious manual work that requires many hours, there are not so many companies in the current ratings. Therefore, the question arose about automating the calculation of the ethics index of companies. To do this, we decided to use non-financial reports of companies from the RSPP website and develop our own algorithm for calculating ESG indicators based on them.

Содержание

1 Введение **4**

2 Цели и задачи проекта **5**

3 План выполнения проекта **5**

3.1 Предобработка корпуса нефинансовых отчетов 5

3.2 Выделение топиков моделью Top2Vec 6

3.3 Сопоставлние топиков выделенных моделью и топиков ESG 7

3.4 Расчет косинусного расстояние между топиками и отчетами 8

3.5 Использование TOPSIS для многофакторного ранжирования 8

3.6 Классификация отчетов 8

4 Результаты исследования **9**

4.1 Сравнение с ESG-рейтингом RAEX 9

4.2 Сравнение с Индексом РСПП «ОТВЕТСТВЕННОСТЬ И ОТКРЫТОСТЬ» 15

5 Заключение **16**

6 Дальнейшие шаги **17**

Список литературы **18**

1 Введение

ESG (от английского Environmental, Social, and Corporate Governance — экология, общество и корпоративное управление) — это концепция, которая включает в себя три аспекта устойчивого развития: забота об окружающей среде, социальная ответственность и корпоративное управление. Есть несколько причин, почему ESG это важно:

- Снижение рисков: ESG-рейтинги помогают компаниям оценить свои риски и возможности, связанные с экологическими, социальными и управленческими аспектами своей деятельности. Это может помочь компаниям принимать более обоснованные решения и снизить риски для себя и для окружающей среды.

- Улучшение репутации: компании, которые придерживаются принципов ESG, могут улучшить свою репутацию и привлечь больше инвесторов и партнеров. Это может привести к увеличению прибыли и росту бизнеса.

- Влияние на общество: ESG-принципы направлены на то, чтобы компании учитывали интересы своих сотрудников, клиентов, партнеров и общества в целом. Соблюдение этих принципов может привести к улучшению качества жизни людей и уменьшению негативного влияния на окружающую среду.

- Повышение устойчивости: ESG-подход помогает компаниям стать более устойчивыми к изменениям в мировой экономике и обществе. Это может снизить зависимость от краткосрочных финансовых результатов и повысить долгосрочную устойчивость бизнеса.

На сегодняшний день рейтинги и индексы этичности ESG формируются различными организациями, такими как рейтинговые агентсва, банки, инвестиционные фонды и другие финансовые институты. Они используют различные методики оценки, учитывают много разных факторы, из-за

чего итоговые результаты не всегда совпадают. Сейчас это делается в ручную, это сложная трудоемкая работа, которая отнимает много времени и ресурсов.

Данный проект призван автоматизировать часть этой работы связанную с выделением фактором из нефинансовых отчетов компаний.

Результатом данной работы стал алгоритм для оценки покрытия нефинансового отчета подтопиками ESG с дальнейшим получением индекса этичности на основе этой оценки. Полученный индекс сравнивался с рейтингами RAEX [4] и РСПП [3].

2 Цели и задачи проекта

Целью данного проекта было получение алгоритма, который автоматически на основе нефинансового отчета компании, оценивает насколько компания соответсвует концепции устойчивого развития ESG. Отмечу, что мы не можем дать полную оценку только на основе нефинансовых отчетов, так как компании не пишут о себе плохо в отчетах. Но мы можем оценить открытость организации и то, насколько она раскрывает данные о своей деятельности.

После этого необходимо сравнить индекс получнный алгоритмом с уже известными рейтингами и проанализировать полученные результаты.

3 План выполнения проекта

3.1 Предобработка корпуса нефинансовых отчетов

В качестве выборки мы использовали корпус нефинансовых отчетов, который есть на сайте Российского союза промышленников и предпринимателей в свободном доступе. Всего там представлено 1388 отчетов разных компаний за разные года с укзанием типа отчета (отчет по устойчивому развитию, социальный отчет, интегрированный отчет, экологический отчёт), сектора, в котором работает компания (образование, здравоохране-

ние, энергетика и т.д.).

После скачивание отчетов в формате PDF с сайта РСПП, они были распарсены по страницам в тексты, были убраны знаки препинания, и оставлены только слова на кириллице. Затем все слова были лемматизированы. После мы посчитали сколько раз каждое слово используется в отчете, и в каком количестве отчетов оно встречается. На основе этой статистики мы убрали слова, которые встречаются слишком редко или слишком часто. В итоге из страниц нефинансовых отчетов получился корпус текстов, которые в дальнейшем использовались в качестве обучающей выборки для модели Top2Vec.

3.2 Выделение топиков моделью Top2Vec

На основе полученного корпуса текстов мы выделили топики, слова сгруппированные по смыслу и теме. Для получения топиков использовался Top2Vec [1] — модель, разработанная для выделения топиков и семантического поиска. Он позволяет обнаруживать группы слов в подаваемом на вход текстовом корпусе, а также генерировать для текстов и слов многомерные векторы. Размер групп слов зависит от настроек алгоритма HDBSCAN. Top2Vec выполняет следующий шаги:

1. Векторизация документов и слов (алгоритмы Doc2Vec, Universal Sentence Encoder, BERT Sentence Transformer)
2. Понижение размерности векторов (алгоритм UMAP)
3. Кластеризация векторов (алгоритм HDBSCAN)
4. Для каждой кластера вычислияется центр тяжести векторов документа в исходном измерении, это вектор темы.
5. N наиболее близких векторов относятся к конкретному топику. N — параметр модели.

Так из полученного корпуса текстов с помощью алгоритма Top2Vec было выделено 324 топика.

Таблица 1: Подтопики, выделенные на основе экспертной оценки

Топик ESG	Подтопики
E	Биоразнообразие
	Вода
	Газ
	Энергия
	Отходы
	Экологический менджмент
	Биоразнообразие
	Климат
S	Трудовые отношения
	Благотворительность
	Безопасность и охрана здоровья
	Оплата труда
	Обучение и развитие
	Профсоюзы и коллективные договоры
	Инвестиции и капитальные вложения
	Отношения с потребителями
	Безопасность продукта
	Налоги
	Отношения с работниками
G	Отчетность и прозрачность
	Управление рисками
	Антикоррупция
	Эффективность и производительность
	Инновации
	Лидерство
	Дивиденды и акционеры

3.3 Сопоставлние топиков выделенных моделью и топиков ESG

Следующая задача, которая перед нами встала — это сопоставление топиков, выделенных моделью с топиками ESG. Для этого была привлечена экспертная оценка[1], на основе которой 25-ти подтопикам ESG были поставлены в соответствие 168 топиков (см. таблицу 1), выделенных моделью Top2Vec, остальные были выкинуты.

[1]Экспертная оценка была проведена Кандидатом экономических наук, доцентом НИУ ВШЭ Сторчевым Максимом Анатольевичем

3.4 Расчет косинусного расстояние между топиками и отчетами

Для того, чтобы определить насколько компания придерживается принципов ESG, нам нужно посмотреть насколько каждый из подтопиков раскрыт в отчете. Для этого закодируем отчеты и множество слов, соответствующее каждому из подтопиков ESG, с помощью TF-IDF (term frequency — отношение числа вхождений слова к общему числу слов в тексте, inverse document frequency — инверсия частоты, с которой слово встречается в корпусе текстов). Затем для всех сочетаний нефинансовый отчет-подтопик ESG посчитаем косинусную близость между нормализованными векторами — мера схожести между текстами.

3.5 Использование TOPSIS для многофакторного ранжирования

После расчет косинусной близости у нас есть 7 факторов для E, 11 факторов для S и 7 факторов для G. Теперь надо определить индекс по каждой букве, чтобы можно было понять какую позицию в рейтинге занимает тот или иной отчет. Для этого используем TOPSIS (The Technique for Order of Preference by Similarity to Ideal Solution) — консенсационный многокритериальный метод анализа решения. В основу этого алгоритма положена следующая идея: выбранная альтернатива должна иметь наибольшее (наименьшее) геометрическое расстояние от идеально-негативного (позитивного) решения. Так TOPSIS позволяет нам посчитать нормированное расстояние до лучшей и худшей альтернативы. В качестве метрики для составления рейтинга возьмем расстояние до худшей альтернативы.

3.6 Классификация отчетов

Альтернативный подход, расмотренный нами — это использование факторов, полученных после расчета косинусной близости для решения задачи классификации. У РСПП есть Индекс «ОТВЕТСТВЕННОСТЬ И ОТ-

КРЫТОСТЬ», который оценивает раскрытие информации компании о себе. Этот индекс ведется с 2017 года. К сожалению, в нем не так много компаний, а тех по которым есть отчеты и того меньше — всего в выборку попало 155 отчетов: 125 — для обучения, 30 — для теста. До 2020 года РСПП присуждал компаниям только две возможные оценки: «A» и «B». Этот рейтинг мы и решили использовать в качестве таргета. Чтобы упростить задачу индекс за 2021 и 2022 год мы также рассматривали как-будто там всего две оценки. Для решения мы выбрали следующий список моделей:

1. CatBoost (CatBoostClassifier)
2. AdaBoost (AdaBoostClassifier)
3. Случаный лес (RandomForestClassifier)
4. Метод опорных векторов (SVC)
5. Логистическую регрессию (LogisticRegressionCV)
6. Перцептрон (MLPClassifier)

4 Результаты исследования

4.1 Сравнение с ESG-рейтингом RAEX

ESG-рейтинг RAEX ведется с 2018 года, в нем представлены позиции в отдельности по каждой букве E, S и G. С каждым годом число компаний, представленных в рейтинге растет. Чтобы посмотреть насколько хорошо наша TOPSIS-оценка совпадает с рейтингом RAEX, мы решили сравнить изменение позиции в рейтингах с течением времени в период с 2018 по 2021 в год. Для этого мы отобрали все компании из ESG-рейтинга RAEX по которым есть нефинансовые отчеты за все четыре года, таких компаний оказалось всего 13. Затем мы пересчитали абсолютные позиции компании в каждом рейтинге в относительные, чтобы позиция компании в конкретный

год была от 1 до 13. После мы перевернули все позиции, чтобы 13 было топом рейтинга.

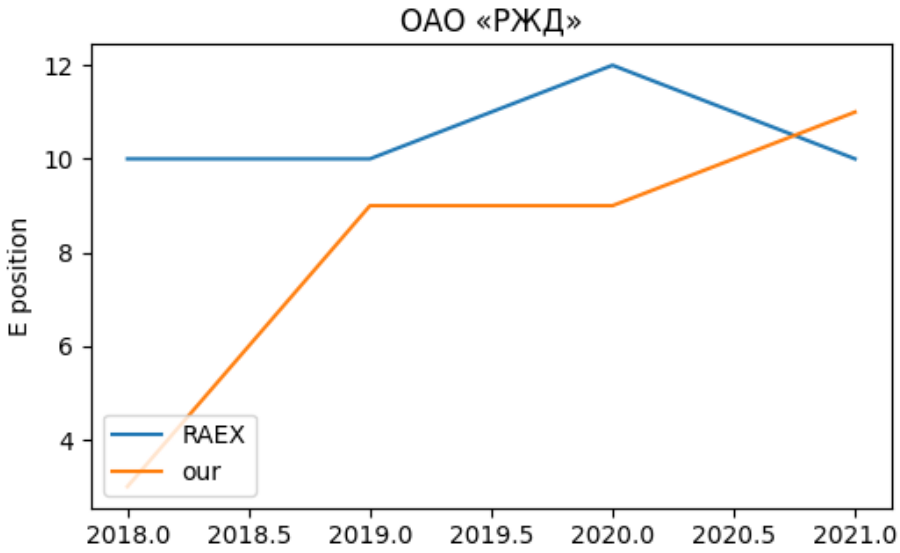

Рис. 1: Сравнение E-рейтингов RAEX и TOPSIS для ОАО «РЖД». На графике видно, что тренды не похожи, но при этом в 2019-ом году и в середине 2020 позиции практически совпали.

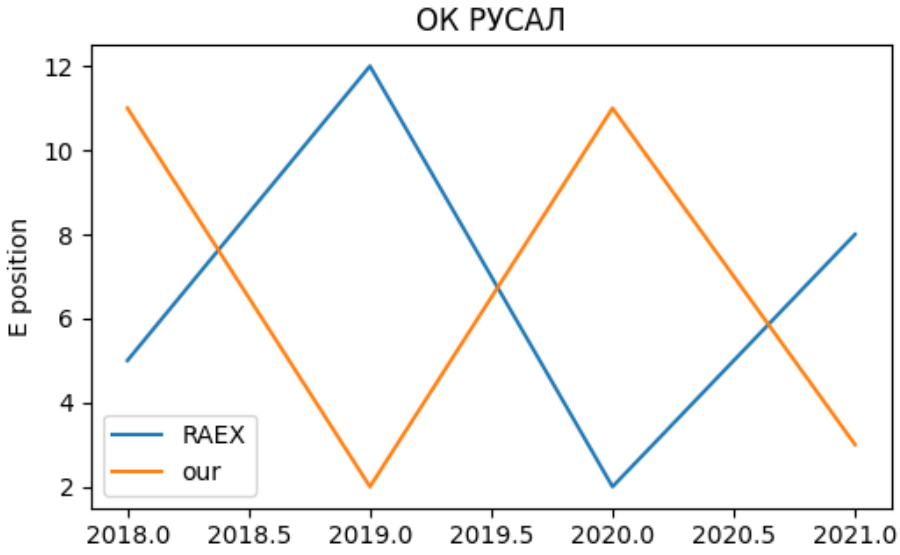

Рис. 2: Сравнение E-рейтингов RAEX и TOPSIS для ОК РУСАЛ. На графике видно, что тренды не похожи. Кажется, что есть смещение в данных, или в отчете РСПП отображена информация за предыдущий год, или рейтинг RAEX смотрит на данные за предыдущий год.

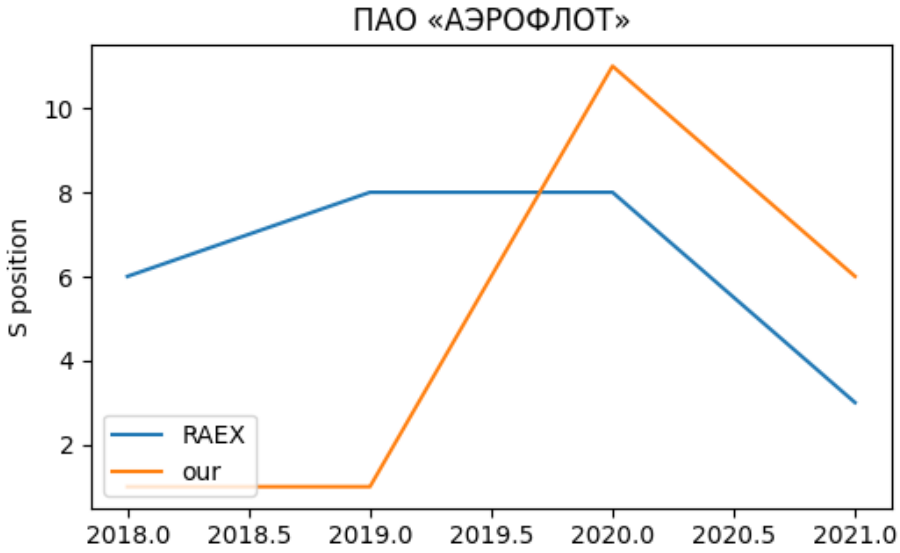

Рис. 3: Сравнение S-рейтингов RAEX и TOPSIS для ПАО «АЭРОФЛОТ». Тренды совпадают частично, в период с 2020-го по 2021 год, при этом абсолютные позиции разные.

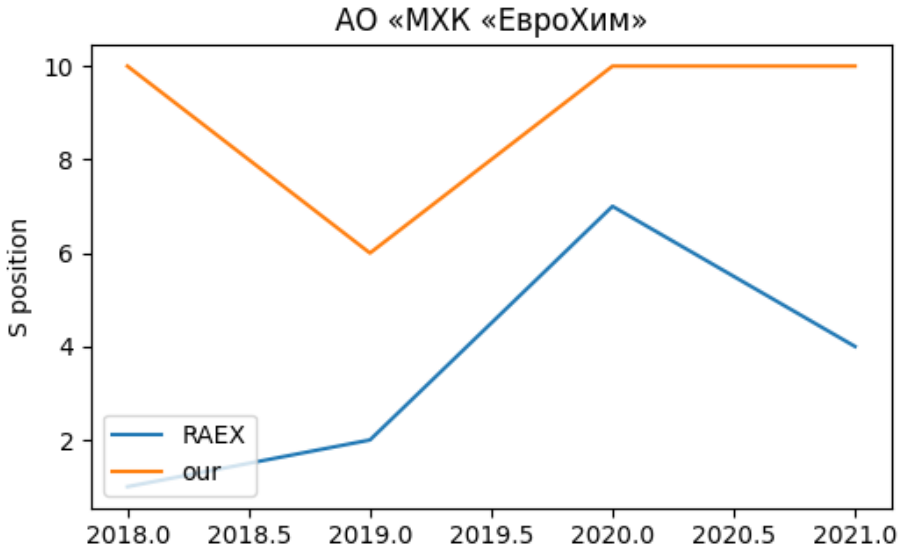

Рис. 4: Сравнение S-рейтингов RAEX и TOPSIS для АО «МХК «ЕвроХим». Тренды совпадают частично, в период с 2019-го по 2020 год, при этом абсолютные позиции разные.

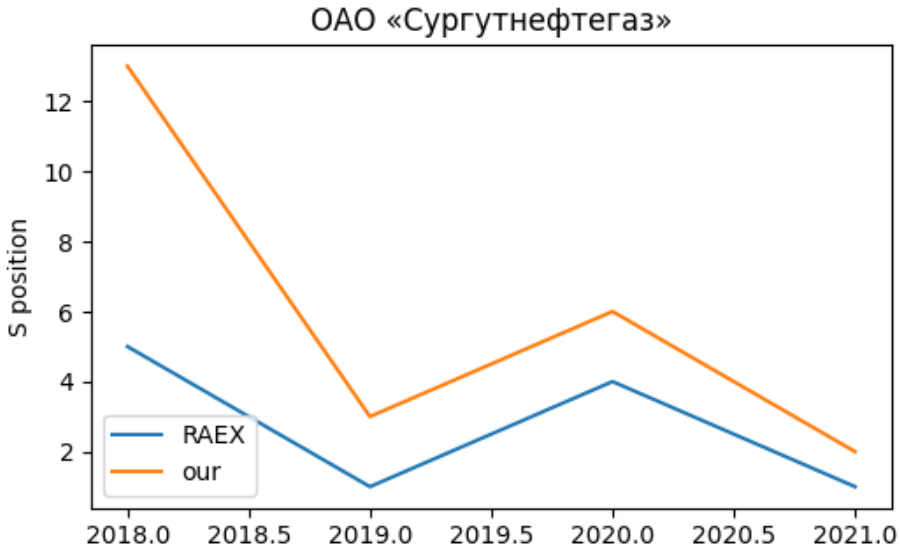

Рис. 5: Сравнение S-рейтингов RAEX и TOPSIS для ОАО «Сургутнефтегаз». Тренды совпадают, при этом абсолютные позиции в каждом из рейтингов разные.

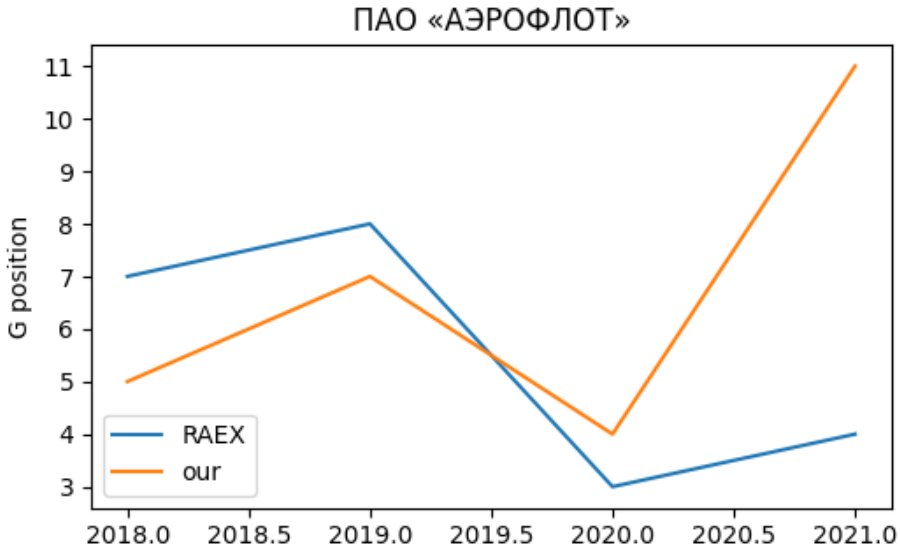

Рис. 6: Сравнение G-рейтингов RAEX и TOPSIS для ПАО «АЭРОФЛОТ». Тренды совпадают, при этом абсолютные позиции в каждом из рейтингов разные.

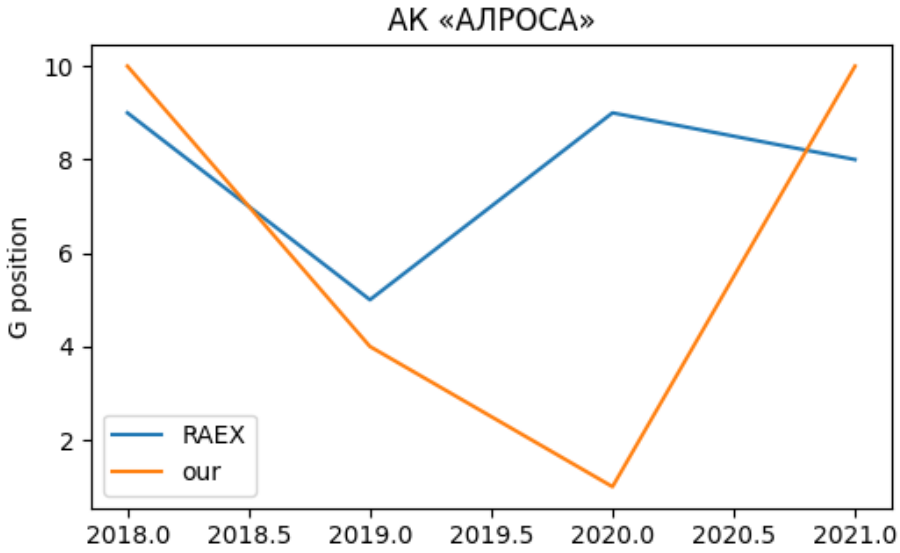

Рис. 7: Сравнение G-рейтингов RAEX и TOPSIS для АК «АЛРОСА». Тренды совпадают частично, возможно, что есть смещение в данных: и у RAEX и TOPSIS есть локальный минимум со сдвигом в один год, в 2019 и в 2020 году соответственно.

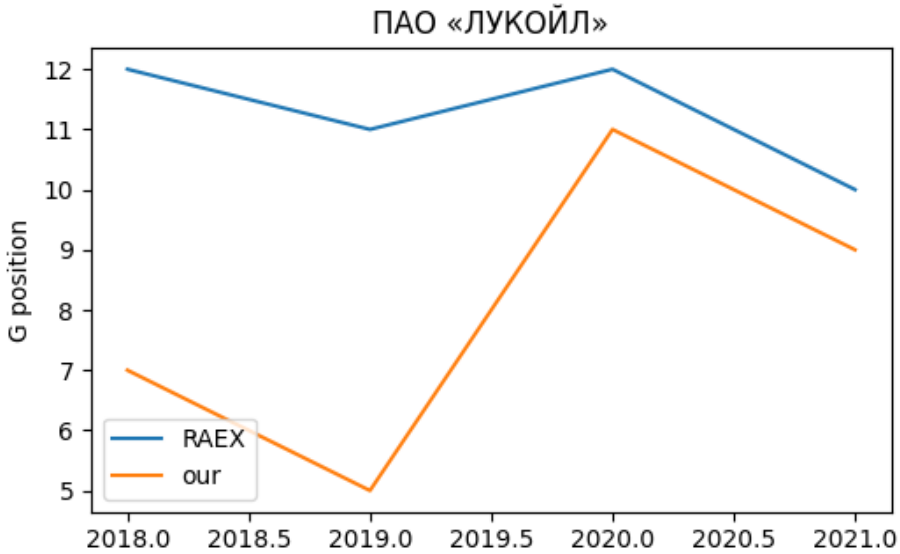

Рис. 8: Сравнение G-рейтингов RAEX и TOPSIS для ПАО «ЛУКОЙЛ». Тренды совпадают, при этом абсолютные позиции в каждом из рейтингов разные.

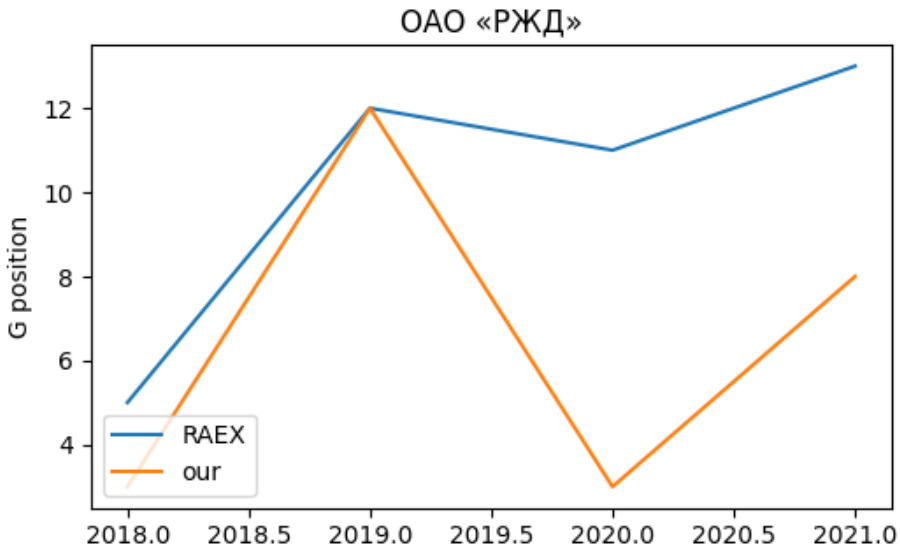

Рис. 9: Сравнение G-рейтингов RAEX и TOPSIS для ОАО «РЖД». Тренды совпадают, при этом абсолютные позиции в каждом из рейтингов разные, есть совпадение только в 2019 году.

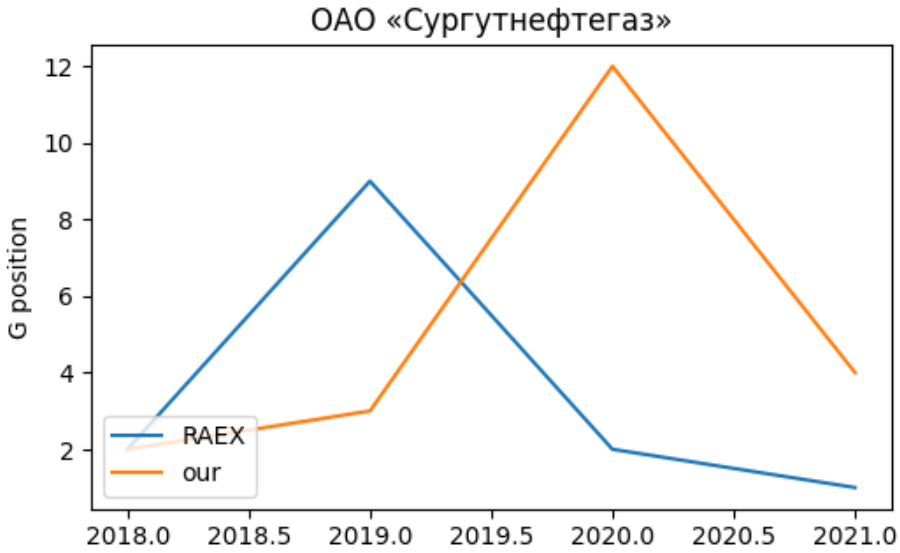

Рис. 10: Сравнение G-рейтингов RAEX и TOPSIS для ОАО «Сургутнефтегаз». Тренды не похожи. Кажется, что есть смещение в данных, или в отчете РСПП отображена информация за предыдущий год, или рейтинг RAEX смотрит на данные за предыдущий год.

Рассмотрим несколько примеров. На графиках (ось ординат — позиция в рейтинге, ось абсцисс — год) представлено изменения позиции компании согласно рейтингу RAEX и рейтингу, полученному нами на основе модели

TOPSIS. На графике 1 тренды не совпадают, при этом рейтинги согласуются по позициям в 2019 году, а также в середине 2020-го года. На графиках 2, 10, 7 тренды не совпадают, при этом кажется, что есть сдвиг в данных, на основе которых составлялся один из рейтингов. Наример, при составлении отчета за 2020 год использовались данные 2019 года. На графиках 5, 6, 8, 9 тренды рейтинга RAEX и полученного нами на основе нефинансовых отчетов совпадат, при этом абсолютные позиции разные. На графиках 3 и 4 тренды свопадают частично, только в определнные периоды.

Также мы проверили есть ли корреляция между рейтингами, корреляция отсутствовала.

Расхождение в данных и остуствие корреляции можно объяснить тем, что рейтинговое агенство RAEX для составления своего ESG-рейтинга использует не только отчеты компаний, а всю доступную информацию — 210 различных индикаторов, из которых 150 общих, остальные зависят от отрасли. Чтобы получить полноценный ESG-рейтинг, аналогичный рейтингу RAEX, к нашим данным нужно добавить информацию из новостей, сайтов компаний и т.д.

4.2 Сравнение с Индексом РСПП «ОТВЕТСТВЕННОСТЬ И ОТКРЫТОСТЬ»

Для сравнения численного индекса, полученного нами на основе модели TOPSIS, с индексом РСПП «ОТВЕТСТВЕННОСТЬ И ОТКРЫТОСТЬ» мы решили посмотреть какой диапазон значений принимает наш индекс для компаний соответстующих каждой букве из индекса «ОТВЕТСТВЕННОСТЬ И ОТКРЫТОСТЬ». Результаты оказались очень плохими: диапазоны для оценок «A» и «B» сильно пересекались, а в некоторые года максимум по знацениям TOPSIS-индекса для худшей оценки «B» даже превышал макимум оценки «A». Данный результат говорит о невозможности сопоставления рейтингов «ОТВЕТСТВЕННОСТЬ И ОТКРЫТОСТЬ» и полученного нами на основе модели TOPSIS.

Также мы пробовали другой подход для сопоставления двух индексов:

Алгоритм	MSE	MAE	ROC-AUC	Accuracy	F1
CatBoost	0.2	0.2	0.89	0.8	0.77
Random Forest	0.17	0.17	0.86	0.83	0.8
AdaBoost Classifier	0.33	0.33	0.67	0.67	0.58
Linear SVM	0.2	0.2	0.79	0.8	0.77
RBF SVM	0.47	0.47	0.10	0.53	0.0
Multilayer perceptron	0.23	0.23	0.76	0.76	0.74
Logistic Regression	0.27	0.27	0.74	0.73	0.73

Таблица 2: Результаты классификации

решение задачи классификации, где признаками выступали факторы полученные нами после расчета косинусной близости, а таргетами были оценки «A» и «B» из индекса РСПП. Результаты классификации предствалены в таблице 2.

Лучше всего себя показала модель CatBoost, но при такой маленькой выборке в первую очередь стоит обратить на более простые модели вроде логистической регрессии и метода опорных векторов, которые тоже неплохо себя показали.

5 Заключение

В результате выполнения данного проекта был получен поэтапный процесс по предобработке нефинансового отчета компании, выделения из него признаков и преобразования их в индекс ESG. На основе данного решения не получится получить полную оценку, так как в нефинансовых отчетах нет большой части информации, которую можно взять из других источников. Но можно использовать данный процесс, как основу для построения полноценного индекса.

Исходный код проекта лежит на платформе GitHub:
https://github.com/MichaelYashchenko/ESG-Project

6 Дальнейшие шаги

Данный проект имеет много направлений для дальнейшего исследования. Рассмотрим некоторые из них:

- Использование всех топиков, выделенных Top2Vec, для классификации, чтобы модели самостоятельно выделяли значимые признаки. Но тут опять встает проблема с размером выборки.

- Использование других методов, помимо косинусной близости, для оценки схожести множества слов.

- Увеличение выборки за счет ручной разметки отчетов. Если собрать больше отчетов и разметить их по одной из методик, можно будет попробовать классифицировать тексты с помощью более мощных моделей, например BERT.

Список литературы

[1] Dimo Angelov. Top2Vec: Distributed Representations of Topics // Computation and Language, August 2020.

[2] Модельная методология ESG-рейтингов // Банк России, Доклад для общественных консультаций, 2023.

[3] Методика составления индексов РСПП по устойчивому развитию (ESG-Индексов), 2022.

[4] Methodology for Assigning Environmental, Social and Governance Ratings to Corporates // RAEX Europe, Frankfurt am Main, September 2019.