Факультет экономических наук

Образовательная программа
«Экономика»

Москва
2022

Использование моделей машинного обучения для предсказания финансовых показателей на основе новостных данных

Драмбян Давид, БЭК186

Руководитель:

доцент факультета экономических наук,

Мамедли Мариам Октаевна

Предпосылки и актуальность

- Гипотеза эффективного рынка — вся существенная информация, находящаяся в открытом доступе, в полной мере отражается в рыночной стоимости активов
- Новости оказывают непосредственное влияние на ожидания и поведение экономических агентов на рынке
- Новостные данные в среднесрочном периоде (дневные колебания) оказывают большее влияние на изменение индекса, нежели тренд или фундаментальная оценка

- Популярность data-driven подхода к инвестированию
- Приложения задачи к реальному рынку

Гипотеза и задачи

Основная гипотеза исследования:

- Согласно установленным предпосылкам, поведение экономических агентов и, вследствие, показатели стоимости рыночных активов, формируемые ожиданиями и действиями этих агентов, могут быть предсказаны на основе данных новостей

Задачи, следующие из нее:

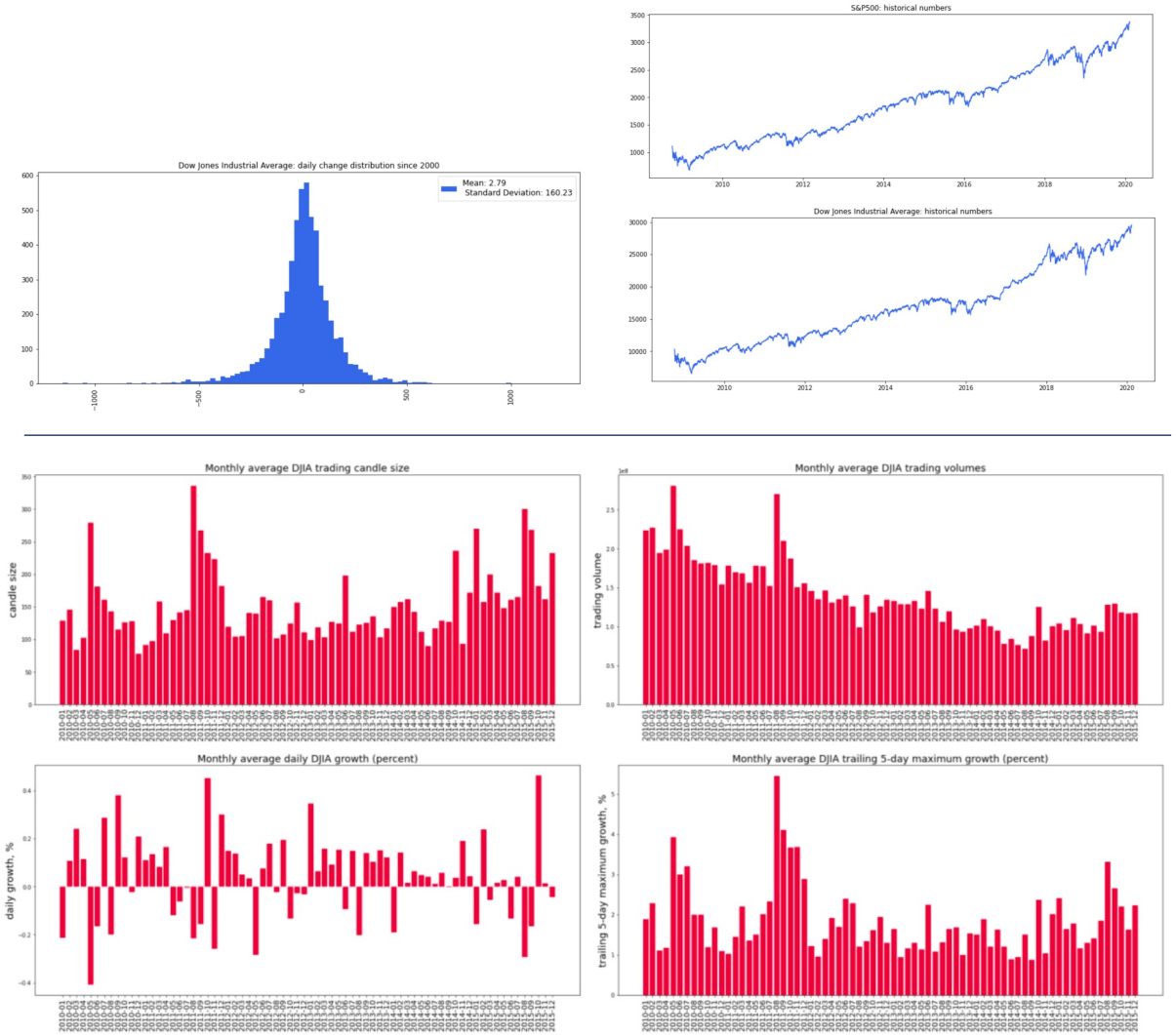
- Эмпирически протестировать гипотезу
- Проанализировать паттерны влияния новостей на изменение котировок
- Оценить качество предсказания моделей машинного обучения, построенных на основе текстов новостей
- Определить наиболее эффективные методы решения проблемы и внести вклад в ее дальнейшее исследование
- Превзойти качество предсказания существующих решений

Обзор литературы — Ключевые моменты

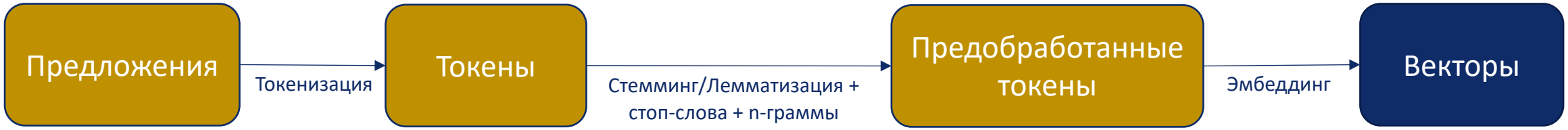
- Использование интерпретируемых вероятностных моделей в сочетании с именами собственными — «Textual analysis of stock market prediction using breaking financial news: The AZFin text system»; Schumaker, Rob & Chen, Hsiu-chin
- Иррациональный дух инвесторов под влиянием новостей оказывает существенное влияние на цены рыночных активов — «Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test»; Lo, Andrew, & Mackinlay, Craig
- Использование индекса негативных слов и финансовых данных — «Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data»; Mao, Huina & Counts, Scott & Bollen, Johan
- Комбинирование технических индикаторов и текстов новостей, использование LSTM-архитектуры — «Deep learning for stock market prediction from financial news articles»; Vargas, Manuel & Lima, Beatriz & Evsukoff, Alexandre

Данные

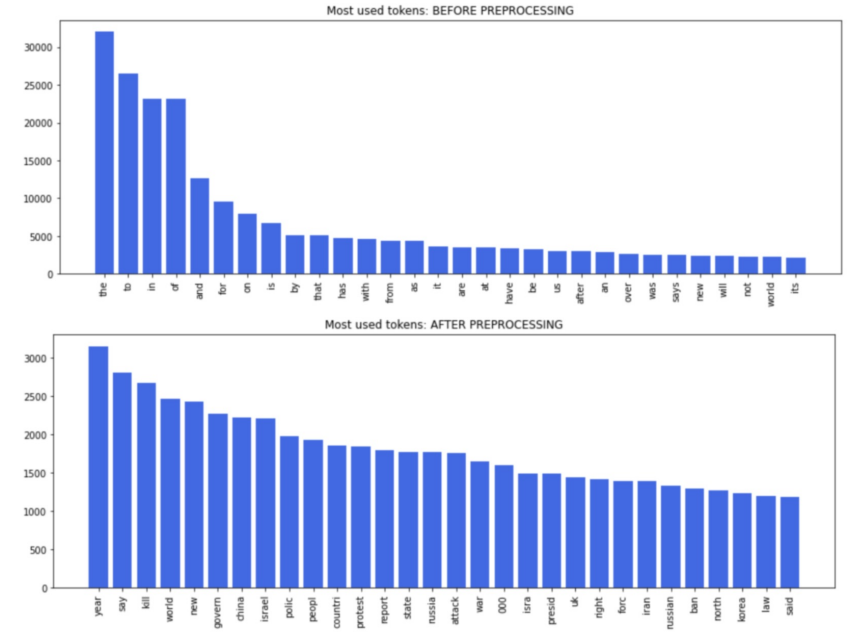
- Целевая переменная: дневное изменение *Dow Jones Industrial Average*
- Текстовые признаки: конкатенация заголовков 25-ти новостей
- Технические показатели: объемы торгов, хронологические показатели, информация о предыдущих торговых сессиях, инструменты технического анализа, …



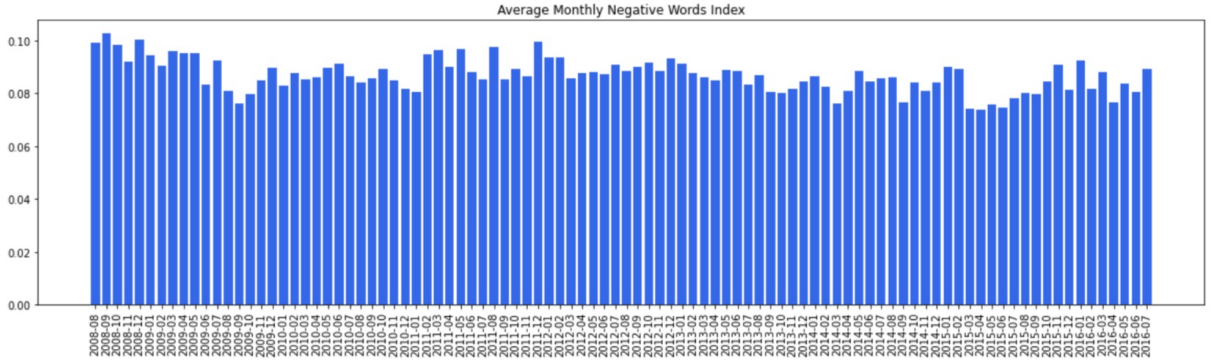
Подготовка текстов



- Предобработка



- Индекс негативных слов



Метрики

- *accuracy* — общая точность модели
- *ROC-AUC*, рассчитанная на основе *True Positive Rate* и *False Positive Rate*

confusion matrix / confusion table	Положительное предсказание	Отрицательное предсказание
Положительный класс	True Positive (TP)	False Negative (FN)
Отрицательный класс	False Positive (FP)	True Negative (TN)

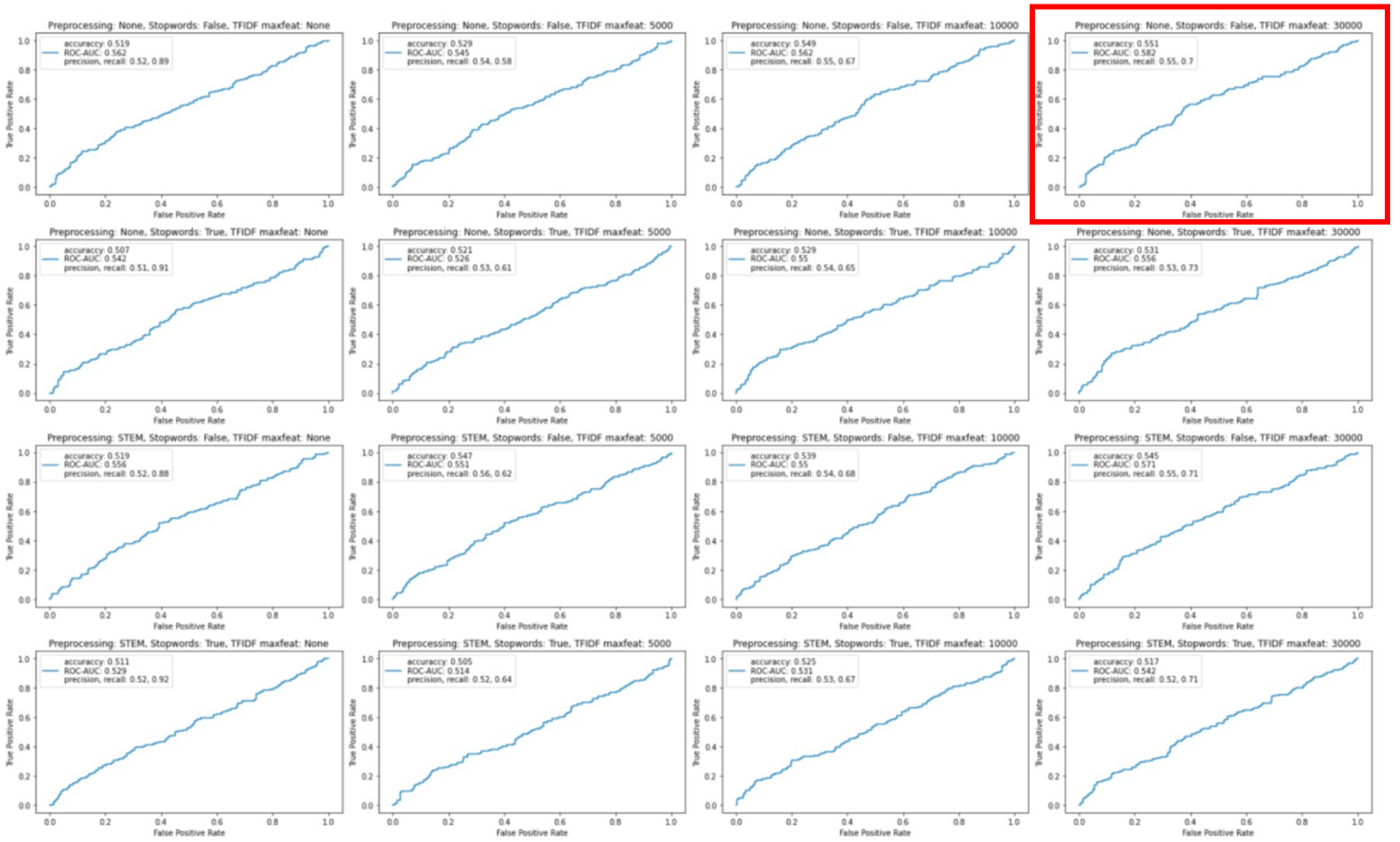
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}, \quad Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(Precision + \beta^2 Recall)}$$

Baseline-модель

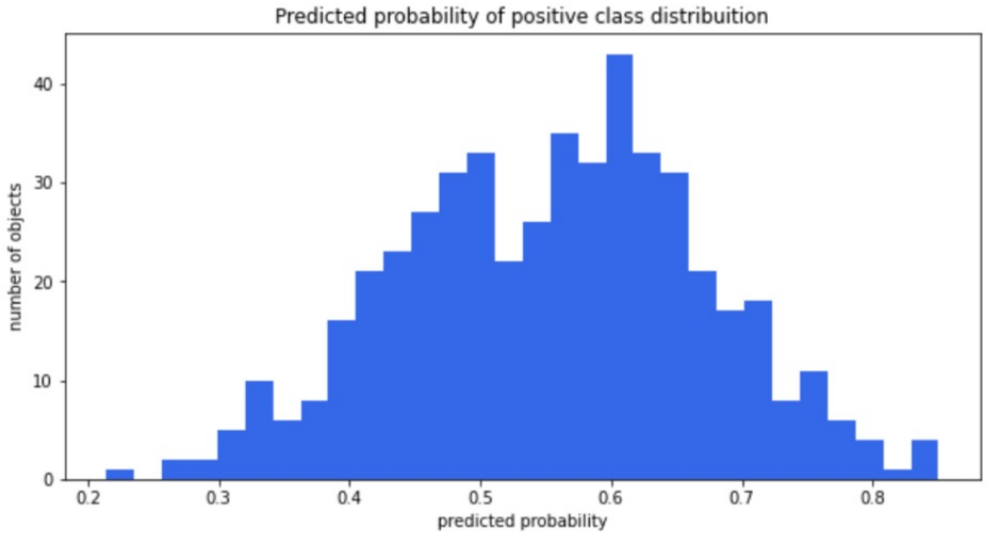
- Эмбеддинг: *TF-IDF*
- Модель: *ансамбль из двух логистических регрессий*
- Лучший accuracy: 0.55
- Лучший ROC-AUC: 0.58
- Precision, positive: 0.55
- Recall, positive: 0.7

Подбор гиперпараметров модели



Baseline-модель
Перемещение трэшхолда вероятности

- Попытка повысить качество модели за счет перемещения трэшхолда вероятности при определении классов



Classification report with probability threshold at 0.5:

	precision	recall	f1-score	support
0	0.55	0.39	0.46	240
1	0.55	0.70	0.62	257
accuracy			0.55	497
macro avg	0.55	0.55	0.54	497
weighted avg	0.55	0.55	0.54	497

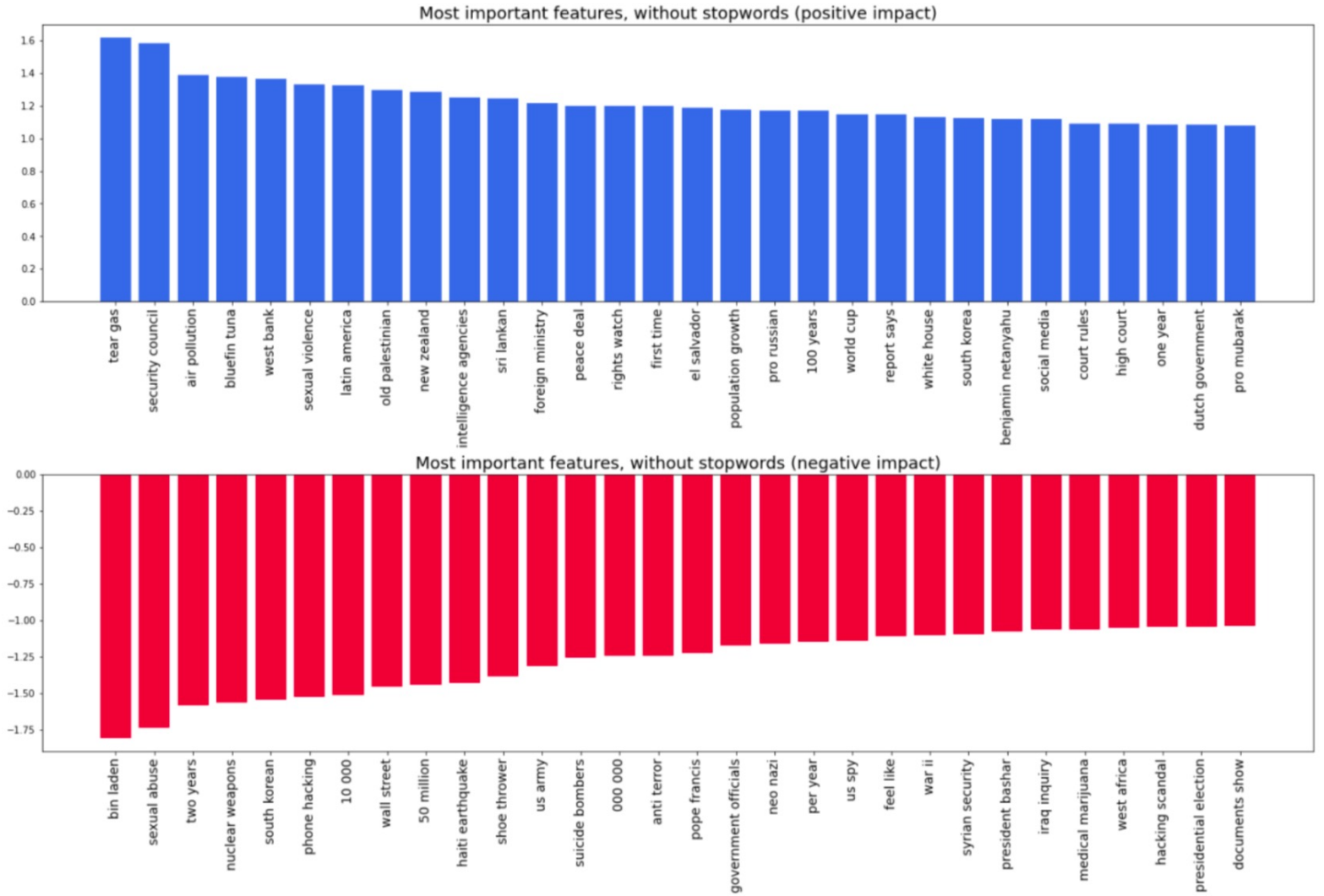
Classification report with probability threshold at distribution mean:

	precision	recall	f1-score	support
0	0.56	0.50	0.53	240
1	0.57	0.63	0.60	257
accuracy			0.57	497
macro avg	0.56	0.56	0.56	497
weighted avg	0.56	0.57	0.56	497

Baseline-модель

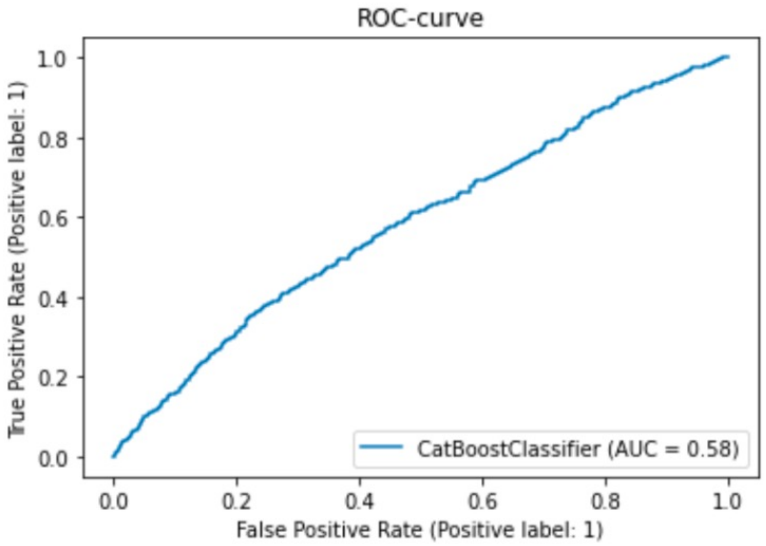
Влияние токенов

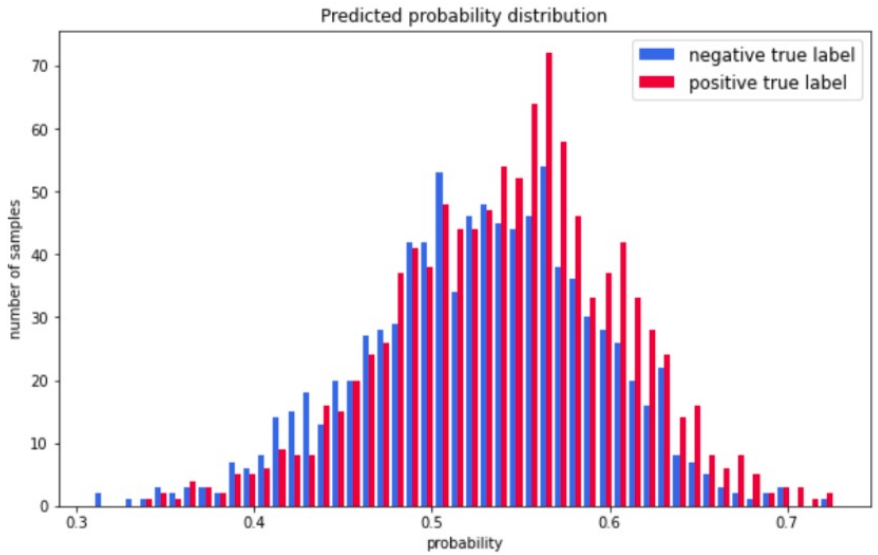
- Нейтральные и несущие мирный или глобализационный посыл биграммы влияют наиболее положительно
- Фразы, содержащие негативный смысл, наиболее отрицательно влияют на таргет



Совершенствование векторных представлений

- Эмбеддинг: *TF-IDF → Word2Vec (подход: количественный → вероятностный)*
- Модель: логистическая регрессия → градиентный бустинг (*catboost*)
- accuracy: 0.56
- ROC-AUC: 0.58





Рекуррентные нейронные сети

- Используем разновидность рекуррентной нейронной сети: *LSTM* (Long short-term memory)
  - cell-state для борьбы с затуханием градиента
- Эмбеддинг: Word2Vec → FinBERT (предобученный)
- accuracy: 0.54
- ROC-AUC: 0.54

Layer (type)	Output Shape	Param #
text_vectorization_1 (TextVectorization)	(None, None)	0
embedding_1 (Embedding)	(None, None, 300)	9063600
lstm_2 (LSTM)	(None, None, 16)	20288
global_max_pooling1d_2 (GlobalMaxPooling1D)	(None, 16)	0
dropout_2 (Dropout)	(None, 16)	0
dense_4 (Dense)	(None, 8)	136
dense_5 (Dense)	(None, 1)	9

Total params: 9,084,033
Trainable params: 20,433
Non-trainable params: 9,063,600

TF-IDF и градиентный бустинг

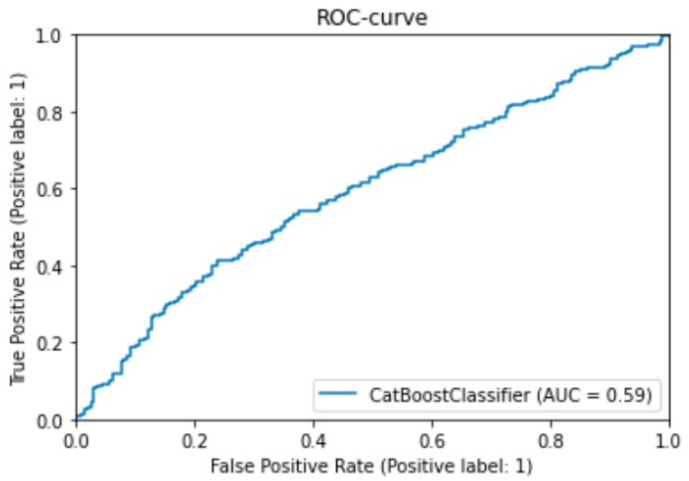
- Попробуем использовать лучшие с точки зрения метрик векторные представления, но повысим сложность модели
- Получаем лучшие результаты

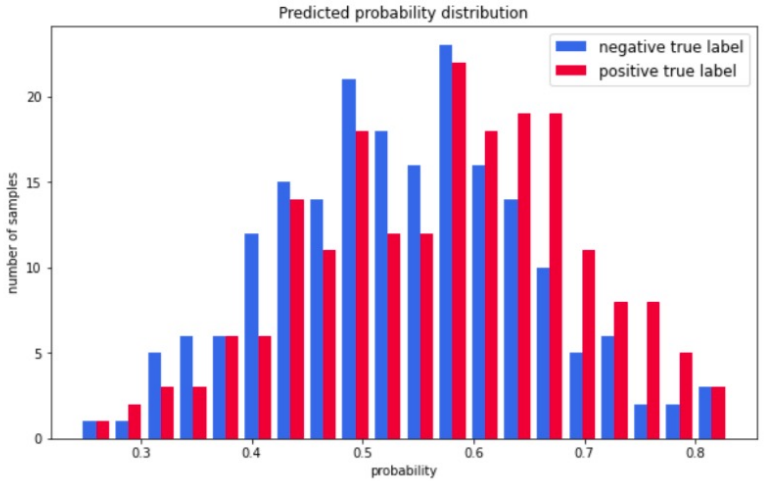
Classification report with probability threshold at 0.5:

	precision	recall	f1-score	support
0	0.55	0.39	0.46	196
1	0.54	0.70	0.61	201
accuracy			0.54	397
macro avg	0.55	0.54	0.53	397
weighted avg	0.55	0.54	0.53	397

Classification report with probability threshold at distribution mean:

	precision	recall	f1-score	support
0	0.57	0.52	0.54	196
1	0.57	0.61	0.59	201
accuracy			0.57	397
macro avg	0.57	0.57	0.57	397
weighted avg	0.57	0.57	0.57	397



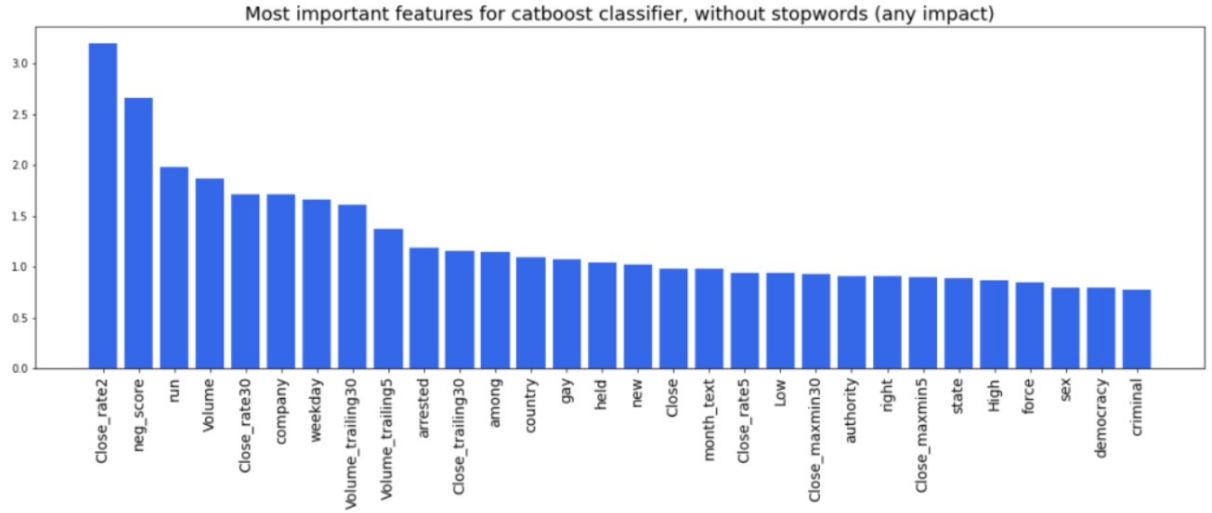


Target encoding токенов

- Эмбеддинг: *TF-IDF → target encoding наиболее популярных токенов*
- *(подход: количественный → привязанный к распределению таргета)*
- Не привел к росту метрик

Classification report with probability threshold at 0.5:

	precision	recall	f1-score	support
0	0.52	0.77	0.62	196
1	0.58	0.31	0.40	201
accuracy			0.54	397
macro avg	0.55	0.54	0.51	397
weighted avg	0.55	0.54	0.51	397



Выводы и дальнейшее исследование

Ключевые выводы:

- показатели стоимости рыночных активов (*DJIA*), действительно, **могут быть предсказаны** на основе текстов новостей (с *ROC-AUC порядка 0.6*)
- семантически **негативные слова** оказывают существенное влияние на изменение котировок
- количественные **методы векторизации** текстов — качественно аппроксимируют общий sentiment в комбинации с градиентным бустингом

Дальнейшее исследование :

- обучение эмбеддингов на собственном корпусе текстов
- более точная связь между новостью и изменением индекса
- поиск оптимальной стратегии выбора слов из новостей перед переходом в векторное пространство

Модель	accuracy	ROC-AUC	F1-score (macro)
TF-IDF + Logistic Regression (Baseline)	0.57	0.58	0,56
Word2Vec + Gradient Boosting	0,56	0,58	0,55
Word2Vec + LSTM	0,55	0,54	0,53
FinBERT + LSTM	0,54	0,54	0,52
TF-IDF + Gradient Boosting	0,57	0,59	0,57
Target encoding + Gradient Boosting	0,54	0,53	0,51

Спасибо за внимание!
Вопросы?