Анализ сентимента телеграм-каналов на российском рынке акций

Корбан Кирилл, Иртуганов Мансур

Student Research Group «Факторный анализ и прогнозирование»
Supervisors: Латыпов Р. Р., Постолит Е. А.

Введение

После ухода нерезидентов с российского рынка акций, по данным Московской биржи, доля физических лиц в объемах торгов на рынке увеличилась с 40% до 80% за период с января по май 2022 года. В это же время усилилась активность в каналах социальной сети Телеграм, посвященных торговле на бирже.

Основная гипотеза

Публикации в телеграм-каналах могут оказывать влияние на поведение агентов на рынке акций, так как доля физических лиц, использующих эти каналы, возрасла.

Задачи исследования

1. Научиться работать с языковыми моделями.
2. Научиться определять, о каких компаниях идет речь в сообщениях.
3. Научиться оценивать сентимент сообщений в отношении выделенных компаний.

Модели

При обработке сообщений используются предобученные модели машинного обучения: BERT [Dev+18], GPT [Bro+20] и RegEx. RegEx использовался для выделения компаний, BERT для определия сентимента, GPT для обеих задач.

Обработка сообщений

Обработка сообщения состоит из:

1. выделения названий в сообщении;
2. фильтрации компаний из списка торгующихся на Московской бирже, устранение неоднозначностей и entity-linking.
3. скоринга сентимента для каждой компании.

Выделение компаний

Требуется найти упоминания компаний в тексте.

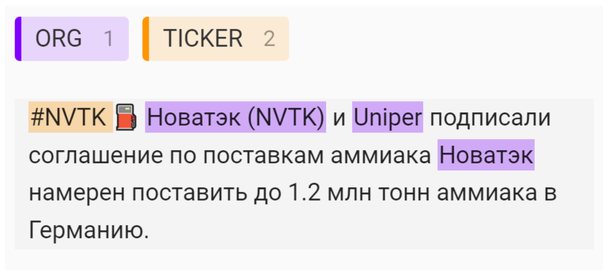

Figure 1: Пример выделения компаний

Использовались 2 подхода:

1. Лемматизировать текст и применить RegEx,
2. Применить BERT-like модель для выделения NER и использовать лемматизацию.

Метод	средний % невыделенных сущностей
RegEx	29.00
BERT-like	53.56

Table 1: Качество выделения компаний.

Оценка сентимента

Для оценки сентимента использовался дообученый на текстах телеграм каналов RuBERT, а так же Catboost на TF-IDF эмбедингах. Разметка состояла из 5 класов. Ниже представленны метрики для сообщений содержащих 1 компанию.

Модель	Accuracy	micro precision	micro F1 score
RuBERT	0.6939	0.7462	0.7107
Catboost	0.6725	0.6872	0.6800

Table 2: Качество сентимента.

Заключение

Мы рассмотрели несколько способов выделения компаний и оценки сентимента. Наилучшим методом выделения компаний оказался RegEX с лемматизацией, так как BERT-like модели выделяют названия компаний с некоторым мусором из-за неформального языка сообщений.
Наилучшим способом определения сентимента оказался RuBERT, но не сильно превзошел catboost.

Список литературы

[Dev+18] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018.

[Rad+18] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[Bro+20] Tom B Brown et al. “Language models are few-shot learners”. In: (2020).