***Проектирование ML-сервиса для прогнозирования котировок акций (для* Advisors' Axiom от Росбанка)**

О проекте
Инвестиционная платформа Advisors' Axiom — это площадка для совместной работы инвесторов сегмента Premium и Private Banking, инвестиционных консультантов и финансовых экспертов, разработанная ПАО РОСБАНК .

Возможности платформы:

ВО ЧТО ИНВЕСТИРОВАТЬ, КОГДА И ЗАЧЕМ	**ТОЛЬКО ВАЖНЫЕ НОВОСТИ**	**РИСК-МЕНЕДЖМЕНТ**
✔️Персонализированные рекомендации ✔️Подборки ценных бумаг ✔️Продуктовый каталог ✔️Модель сбалансированного инвестиционного портфеля	Показывает только те новости, которые могут повлиять на цену ваших активов, и отмечает тональность этого события	Оценка риска портфеля и инвестиционное профилирование

Для проекта была разработана нейросеть для прогнозирования котировок акций в зависимости от тональности финансовых новостей. Клиент получает подборку новостей, связанных только с активами в его портфеле. Полученная информация поможет клиентам, инвестиционным консультантам и финансовым экспертам более эффективно инвестировать денежные средства в ценные бумаги компаний. Как известно, негативные новости чаще всего приводят к снижению стоимости акций, позитивные - наоборот.

Расскажем подробнее о разработке этой нейросети.

Задача

Задача сервиса — получение, оценка (классификация) финансово-экономических новостей на основе машинного обучения модели нейронной сети для классификации текстов финансово-экономической направленности по трем видам тональностей:

- позитивная,
- нейтральная,
- негативная,

и генерация на их основе кратких анонсов на русском языке.

Реализация

Генерация анонсов

Для генерации на основе классификации текстов кратких анонсов на русском языке (с возможностью привязки к конкретной компании) используются две нейросети. Одна генерирует анонс из нескольких слов, которые мы используем в качестве заголовка. Вторая — более длинный текст, который мы используем в качестве анонса.

Все на английском языке. На русский переводится с помощью API Яндекс.Переводчик.

Привязка новости (а значит, и анонса) к компании производится поиском в заголовке новости упоминания компании.

Создание модели

Обучена модель, классифицирующая англоязычные финансовые новости на позитивные и негативные. Определяет позитивную или негативную тональность финансовой новости из любого англоязычного источника (FinViz, например).

Модель выдает уровень уверенности в своей оценке. Предлагается использовать только классификации с высоким уровнем уверенности, а остальные новости помечать как нейтральные.

Для обучения модели потребовалось минимум 10 тысяч размеченных финансовых новостей из наших источников. Чем новости свежее, тем лучше. Их лучше разметить с помощью финансового индикатора, подсказанного финансовым аналитиком.

Репозиторий не содержит код для обучения моделей. Содержит только для генерации анонсов с помощью любой из 4 моделей. Обучение своей модели потребует несколько суток работы видеокарты, поэтому лучше его избежать.

Датасеты

1. newsroom — миллион новостей с анонсами, без категорий.
2. multi_news — 50 тысяч новостей с анонсами, без категорий.
3. gigaword — 4 миллиона статей с анонсами в одно предложение.
4. cnn_dailymail — статьи с CNN и Daily Mail. Анонсы состоят в основном из предложений статьи, поэтому и сгенерированные анонсы зачастую состоят из предложений тестовой новости. Анонсы длиной в 2–3 предложения.

Данные для обучения модели

Исходный датасет содержит 22 297 размеченных новостей Reuters за 2006–2015 годы. Он взят из репозитория к научной статье Learning Target-Specific

Representations of Financial News Documents For Cumulative Abnormal Return Prediction.

Для разметки новостей Reuters использовался финансовый индикатор Cumulative Abnormal Return Prediction. Для компании из новости вычисляется равновзвешенный рыночный индекс, включающий разницу акций на NYSE, Amex, NASDAQ между днем до дня публикации новости и днем после. Если торговый день закончился к моменту публикации новости, то этот индекс вычисляется на день позже.

Более детальное описание из статьи Learning Target-Specific Representations of Financial News Documents For Cumulative Abnormal Return Prediction:

Cumulative Abnormal Return The task that we attack in this paper is *Cumulative Abnormal Return Prediction (**CAR**)*. Formally, the abnormal return AR_{jt} of a firm j on a date t is the difference between its actual return R_{jt} and the expected return $\hat{\mathrm{R}}_{jt}$, $\mathrm{AR}_{jt} = \mathrm{R}_{jt} - \hat{\mathrm{R}}_{jt}$. The expected return $\hat{\mathrm{R}}_{jt}$ can be estimated by an asset price model based on historical prices, or approximated by the market return in a short-term event window (Kothari and Warner, 2004). The cumulative abnormal return CAR_j of the firm j in an n-day time window is calculated by summing up the daily abnormal returns in the period (Eq 1). In this paper, we adopt the commonly used three-day window (-1,0,1), which we denote as CAR_3 and day 0 is the day when the current news documents are released.

$$\mathrm{CAR}_j = \sum_{t=1}^{n} \mathrm{AR}_{jt} \quad (1)$$

We collect publicly available financial news articles from Reuters from October 2006 to December 2015. In our preliminary experiments, we find that a news document is more likely to be relevant to a firm only if it is mentioned in the news abstract. Thus, we only include news documents mentioned at least one public listed firm in the U.S. security market. We group the news documents per firm per event date. If the news is released during a trading hour, day 0 is the current day, otherwise day 0 is the next trading day. We compute the expected return $\hat{\mathrm{R}}_{jt}$ by the return of equally-weighted market index including all the stocks on NYSE, Amex, NASDAQ.

Использование аннотации (анонса) новости вместо ее текста показало плохие результаты.

Протестировано обучение модели на отзывах с Yelp. Они короче, разнообразие позитивных и негативных формулировок намного ниже, поэтому модель получается точнее, чем на финансовых новостях.

Подготовка данных

Настрой новости

Для усиления корреляции между настроем новости и соответствующим ей финансовым показателем решено использовать для обучения модели только самые позитивные и негативные новости. В качестве позитивных новостей взяты новости с индикатором более 0.05. В качестве негативных новостей взяты новости с индикатором менее –0.05.

Также были протестированы пороги в 0.03, 0.07 и 0.1. Они дали меньшую точность модели, так как ей было сложнее найти корреляцию между настроем новости и соответствующим ей финансовым показателем.

Порог	Число новостей	Точность
0.1	2227	71%
0.07	3456	72%
0.05	5187	73%
0.03	8627	71%

Нормализация текста

Тексты новостей уже были частично подготовлены в исходном тексте: знаки препинания окружены пробелами, тексты переведены в нижний регистр.

Затем в текстовом редакторе удалены теги параграфа и новой строки и другие лишние элементы.

Удаление стоп-слов показало ухудшение модели.

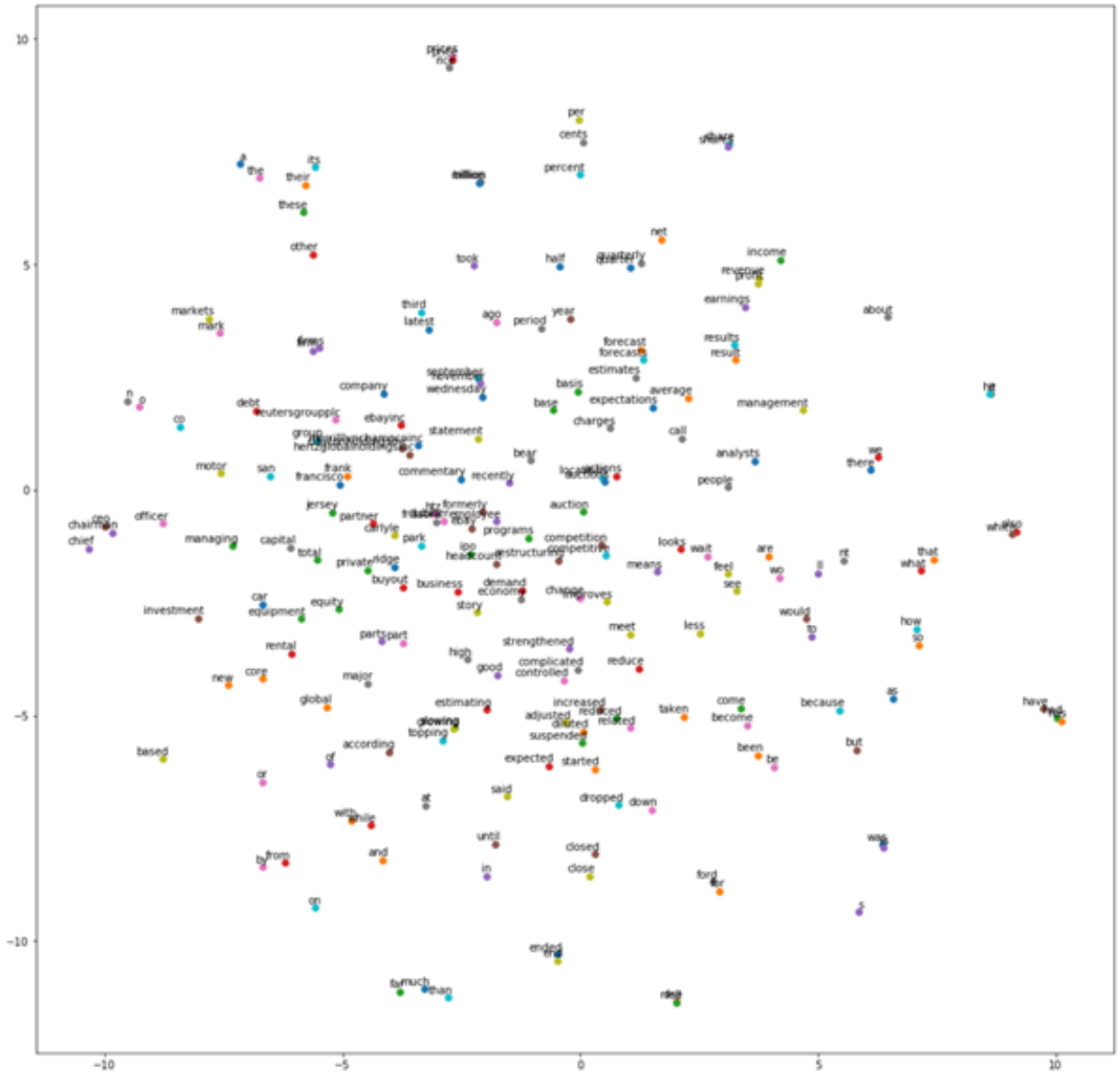
Протестированы различные варианты автоматической подготовки текстов. Лучший вариант находится в скрипте обучения модели.

Обучение модели

Обучение на видеокарте RTX 2070 S выполняется за 1–2 часа.

Обучение дескриптора слов

Обучение своего дескриптора слов (word embedding) дало более высокую точность, чем использование предобученного на датасете wiki-news, так как в финансовых новостях много специализированных терминов.



Кликните на изображение, чтобы увеличить его

Подбор параметров обучения модели

Используется LSTM. Код для обучения модели содержит вариант использования GRU вместо LSTM, но он не работает на версиях TensorFlow 1.13, 1.15, 2.0. Предположительно, GRU будет работать на старой версии 1.10.

Протестированы learning rate 0.0005, 0.0007, 0.001, 0.002. 0.001 дает наивысшую точность.

Протестированы batch size 32, 64, 128. 64 дает наивысшую точность.

Протестированы ограничения количества предложений в новости 15, 30, 45, 50 и 55. Ограничение в 55 предложений показало наивысшую точность. Большее максимальное количество предложений требует больше видеопамяти.

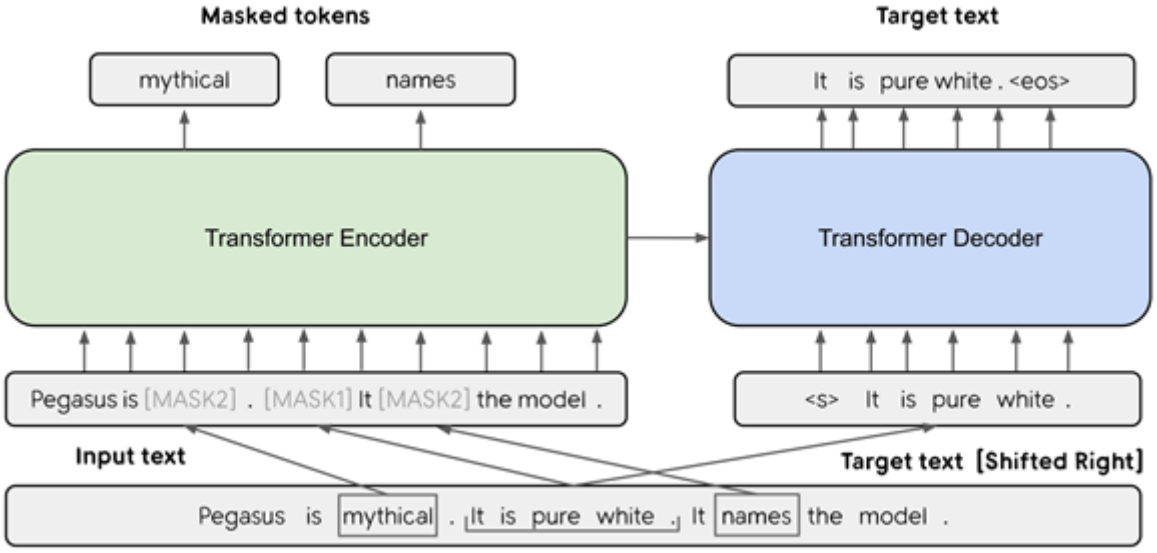
Протестированы ограничения количества слов в предложении 50 и 80. Ограничение в 50 слов показало наивысшую точность.

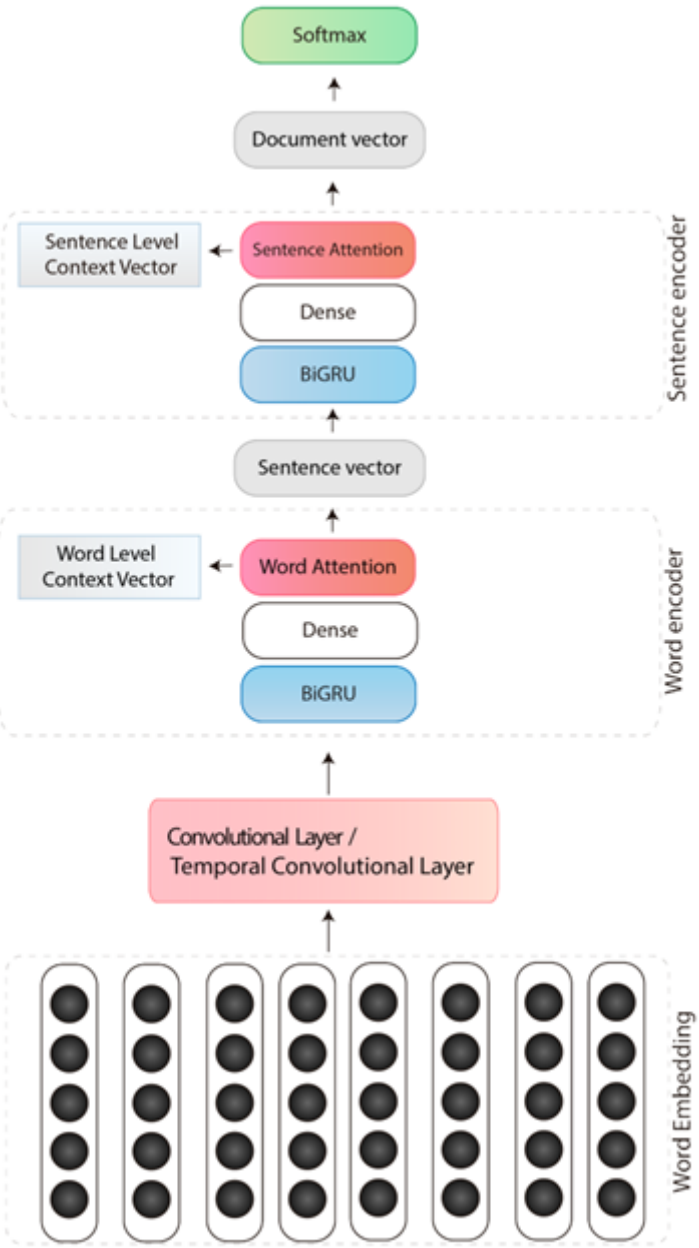
Оптимальная длительность обучения — 6 циклов.

Используемая архитектура

Архитектура модели описана в научной статье Hierarchical Attentional Hybrid Neural Networks for Document Classification. Она реализована в репозитории к этой статье.

В модели используются convolutional neural networks, LSTM, and attention mechanisms.





Проигравшая архитектура

Другая протестированная архитектура нейронки описана в научной статье [Learning Target-Specific Representations of Financial News Documents For Cumulative Abnormal Return Prediction]. [Репозиторий] этой статьи выглядел наиболее подходящим для классификации англоязычных финансовых новостей о любых компаниях. Но точность модели (мера F1) оказалась лишь 60%. И не удалось использовать модель для классификации одиночной новости. Удается запускать только тест на файле с 2000 новостями:

python main.py --gpu 0 --model TE --resume model/TE_avg_16_100_100_0.0005_0.1_epoch_10_17974_model_136_microf1=0.641.pth

Репозитории с альтернативными архитектурами

1. An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation — архитектура и предобученные модели для английского языка.
2. Bart — архитектура и предобученные модели для английского языка и для русского.
3. summarus — предобученные модели для русского языка.

Тестирование на современных новостях с finviz

Новость	Классификация	Уверенность
LG Chem	neg	58%
Daimler	pos	73%
Foxconn	neg	60%
Big Hit	neg	78%
Schlumberger	neg	85%
LVMH	pos	64%

Aligos	pos	54%
British Airways	pos	69%
Vans	pos	73%
Pret A Manger	neg	56%

Похоже, что модель классифицирует реальные новости не лучше подброшенной монетки и обращает внимание не на те слова. Для обучения модели требуется более крупный и современный датасет.

Оценка скорости модели и требования к серверу

Модель загружается один раз, при запуске контейнера, за 1–2 секунды.

Классификация одной новости выполняется на процессоре AMD Ryzen 9 3900X за 1–2 секунды.

После запуска сервис занимает 1 гигабайт оперативной памяти.

Файлы контейнера с библиотеками и моделями занимают на диске 1–2 гигабайта. Большую часть этого пространства занимает TensorFlow.

Итог работы

Приложение с заданным периодом времени автоматически считывает свежие новости из определенного перечня сайтов-источников, умеет получать текст новости по ссылке (URL) на страницу сайта-источника.

Каждая новость, попадающая на оценку и формирование анонса, автоматически привязана по заданному перечню тегов к необходимой компании.

Изображения подбираются для новости автоматически на основе тегов категории компании.

Для перевода новости на русский язык используются внешние сервисы, обеспечивающие качественный машинный перевод текста.

Корректировка анонсов, изображений и публикаций выполняется через панель администрирования модератором контента.