Применение методов оценки тональности текстов для предсказания поведения финансовых рынков

СУХАРЕВ ПАВЕЛ СЕРГЕЕВИЧ

НАУЧНЫЙ РУКОВОДИТЕЛЬ: ДОЦЕНТ, К.Ф.-М.Н. ГРИГОРЬЕВ Д. А

КАФЕДРА ИНФОРМАТИКИ СПБГУ

Существующие подходы

Прогнозирование фондового рынка основанное на исторических ценах акций

Прогнозирование на основе анализа настроений в новостях

Прогнозирование на основе анализа настроений в социальных сетях в т.ч. Твиттере

2

Прогнозирование фондового рынка – актуальная задача и по сей день. Вложение капитала в ценные бумаги все чаще рассматривается инвесторами как наиболее привлекательная альтернатива. Изучение и разработка методов прогнозирования стоимости различных ценных бумаг представляет большой практический интерес для инвесторов. Существует несколько подходов для прогнозирования фондового рынка. Так как гипотеза эффективного рынка говорит, что движения финансового рынка зависят от новостей, текущих событий и выпусков продуктов, то подход прогнозирования на основе настроений пользователей в социальных сетях вызывает особый интерес. В качестве изучаемой социальной сети был выбран Twitter.

Фундаментальные показатели

Основные показатели, использующиеся при фундаментальном анализе компании – общая выручка, чистая прибыль, EBITDA, активы и обязательства, рыночная капитализация, а также коэффициенты P/E и P/S

коэффициент P/E = цена/прибыль на акцию

3

Для принятия взвешенного и рационального решения инвестору доступны две группы методов: методы фундаментального анализа и методы технического анализа. Фундаментальный анализ, как правило, применяется для изучения финансово–экономического состояния компании и позволяет ответить на два основных вопроса: акции какого эмитента могут принести наибольший доход, и какова «справедливая» (внутренняя) цена рассматриваемой акции. При этом фундаментальный анализ абстрагируется от колебаний котировки акции на рынке. В качестве исследуемого фундаментального показателя был выбран индекс P/E. Данный коэффициент показывает соотношение между ценой акции и прибылью на акцию компании. Он сообщает степень уверенности инвесторов в будущем компании.

Цель

Исследование возможности применения методов анализа тональности текстов для твиттов и применение их для прогнозирования коэффициента P/E

4

На данном слайде представлена цель бакалаврской работы. Она заключается в исследовании возможности предсказывать индекс P/E, с помощью анализа тональности твиттов из Твиттера.

Задачи

- Сбор данных из твиттера для определенных компаний
- Изучение основных методов анализа тональности текстов
- Их применение к данной задаче
- Анализ полученных результатов

5

Задачи, которые были поставлены в ходе данной работы. Для начала нужно собрать данные, на которых мы будем проводить исследования. Затем изучить различные современные подходы и методы для анализа тональности текстов и возможность их применения к данной задаче. Проанализировать результаты и выявить возможность строить предсказательную модель.

Сбор данных

Twitter API

Get Old Tweets Programatically

Macrotrends

6

Для исследования требовалось собрать набор данных для каждой компании, который состоит из твиттов и различных фундаментальных показателей. Twitter предоставляет доступ к данным посредством Twitter API, но в доступе к нему отказали, поэтому пришлось использовать альтернативные способы. В данном исследовании, твиты собирались с помощью “Get Old Tweets Programatically”. Фундаментальные показатели, в том числе индекс P/E, были взяты с сайта Macrotrends.

Размер входных данных

Microsoft 2.500.000 твиттов

Apple 3.100.000 твиттов

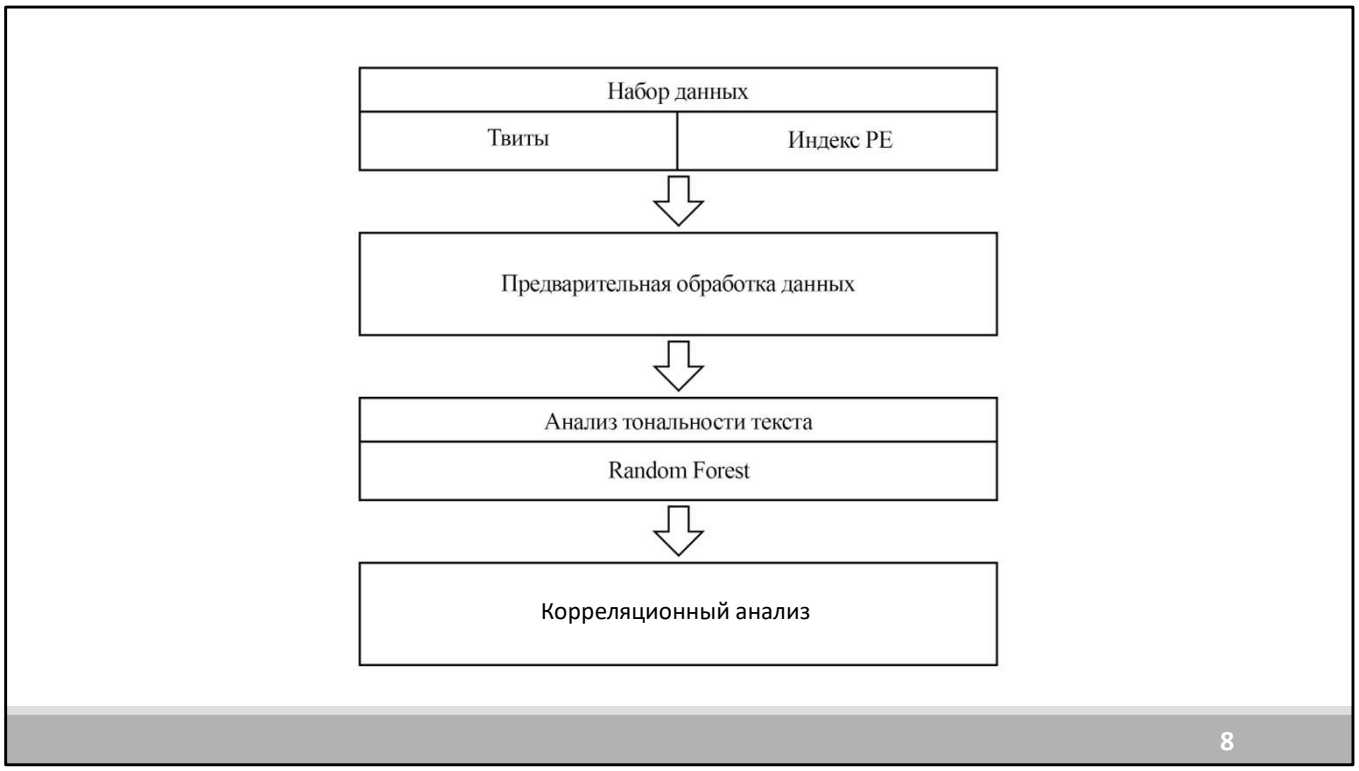
Disney 2.400.000 твиттов

Amazon 2.900.000 твиттов

с 31 ноября 2006 года по 31 декабря 2018

7

Были выбраны 4 компании, и для каждой был собран значительный объём твитов.



На слайде показана модель анализа тональности твитов.

Предварительная обработка данных

токенизация

удаление стоп-слов

сопоставление регулярных выражений для специальных символов

стемминг

9

Для того чтобы алгоритмы работали корректно, твиты нуждаются в предварительной обработке, которая состоит из нескольких этапов. С помощью токенизации, твиты разделены на отдельные слова на основе пробела. Формируем список отдельных слов для каждого твита. Удаляем стоп-слова. С помощью регулярных выражений заменяем URL-адреса, хэштеги и упонимания пользователей соотвественно на URL, HASHTAG, USER. И дальше приводим все однокоренные слова к единому виду.

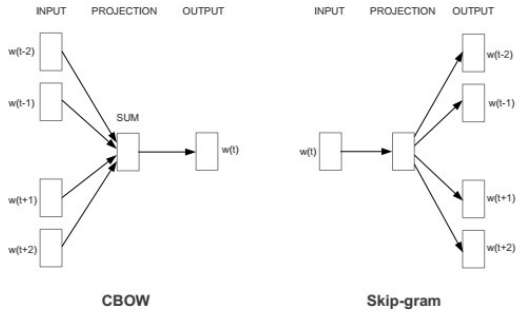
Анализ тональности текста

Word2Vec

Skip-gram

Размер окна 10

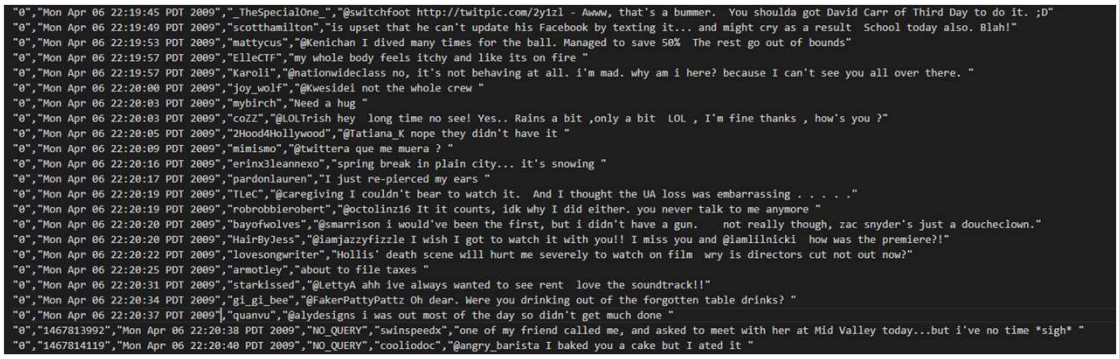
Размер вектора - 300



10

Для анализа тональности текстов был выбран набор инструментов Word2vec, который реализует две основные архитектуры — Continuous Bag of Words (CBOW) и Skip-gram. Принцип работы CBOW — предсказывание слова по имеющемуся контексту, а Skip-gram наоборот — для данного слова предсказывается контекст — слова, стоящие до и после данного слова. Была выбрана конфигурация Word2Vec с Skip-gram т.к. он работает более точно, чем CBOW, особенно для относительно редких слов. На основе проведенных экспериментов, размер окна был взять равным 10, а размер вектора – 300.

Обучение модели



11

В качестве обучающего набора данных использовался уже готовый датасет из твитов собранный для Sentiment analysis, который содержит примерно 90000 твитов, имеющих положительную или отрицательную окраску. Данный набор также был дополнен приблизительно 1 тыс. твитов по компаниям, которые участвуют в исследовании.

Классификация текстов

Random Forest

scikit-learn

Machine Learning Algorithm	Word2Vec			
	Accuracy	Precision	Recall	F-Measure
Random Forest	81,23%	0,749	0,788	0,77

12

В качестве алгоритма классификаций в этом эксперименте используется Random Forest. Алгоритм уже реализован в пакете scikit-learn, и все, что нам остается это вскормить наши текстовые данные и указать количество деревьев. Дальше алгоритм тренируется на обучающей выборке и сохраняет все необходимые данные.

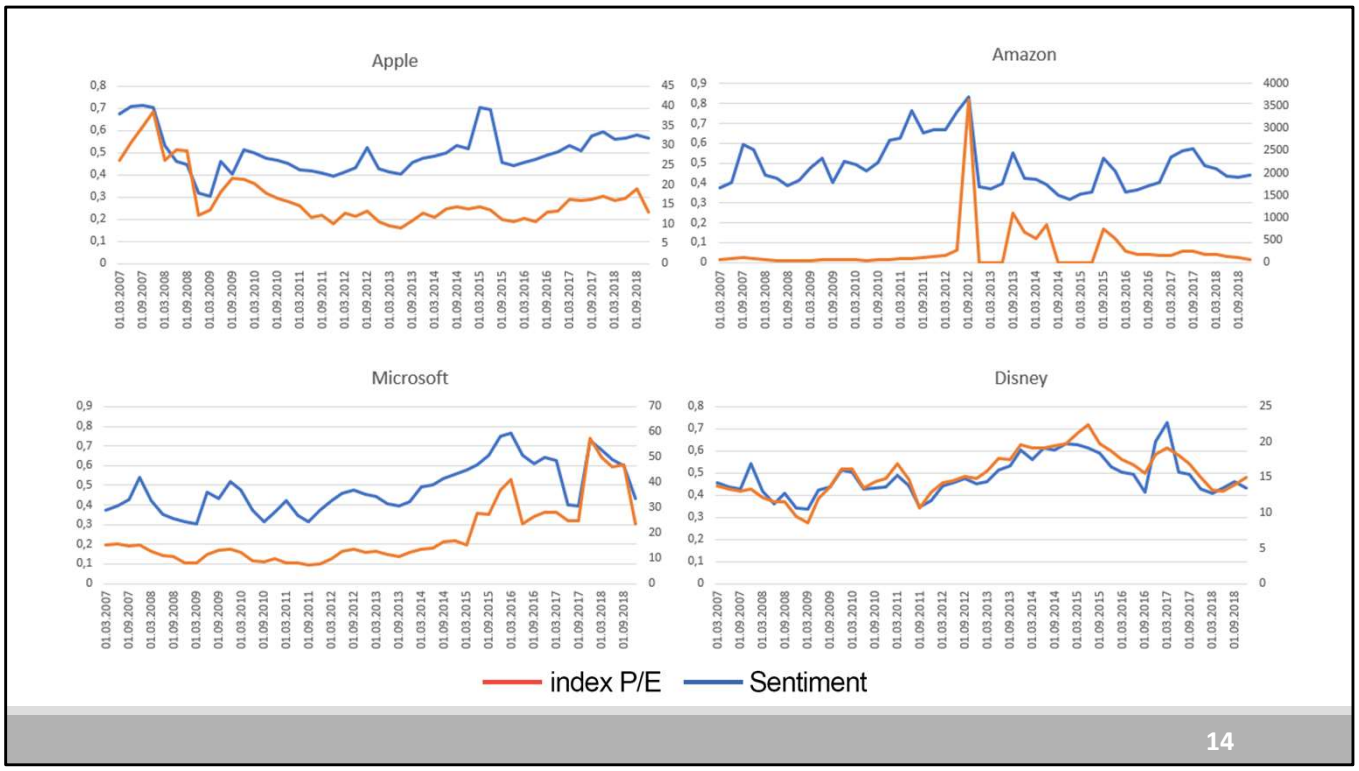
Инфраструктура

Xiaomi Laptop
- Intel(R) Core(TM) i5-7300HQ 2.50GHz
- 8 Gb RAM

Google Colab Laboratory
- T4 or P100 GPU
- 25 GB RAM
- 107 GB DISK

13

Для сбора такого значительного объема данных и проведения анализа полученных результатов быо достаточно мощности собственного ноутбука, так как эти процессы не требуют большой вычислительной мощности. Но сбор твитов всё-таки он занял довольно продолжительное время, в связи с плохой оптимизацией приложения “Get Old Tweets Programatically”. Обучение производилось на платформе Google Colab. Вычислительной мощности машин вполне хватает, для проведения такого рода исследований.



Такие результаты мы получили для каждой компании. Мы можем видеть, что P/E Amazon в 2012, 2013 - 2014 был очень высок. Это говорит о том, что рынок ожидал высокий рост прибыли в будущем. Amazon оценивали как технологическую компанию (с высоким потенциалом будущих доходов от продуктов / услуг с высокой маржой) на более низких доходах Amazon, как розничной компании (низкая маржа, высокая выручка от розничных продаж). В промежутки с 31.12.2012 – 30.06.2013 и с 30.09.2014 – 30.06.2015 мы можем наблюдать значение индекса в нуле. В эти периоды прибыль на акцию (EPS) у Amazon была отрицательная, и индекс не давал никакой информации. Рынки не всегда настолько эффективны и рациональны, как это предполагается, и могут застрять в иррациональном состоянии.

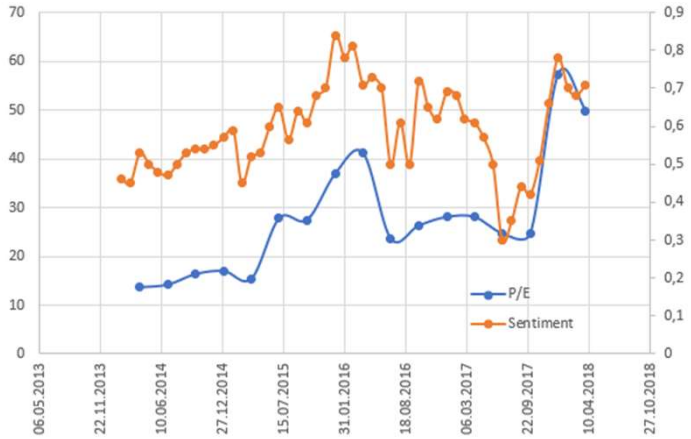
Статистический анализ полученных результатов

Компания	Корреляция с индексом P/E	Корреляция с E	Корреляция с ценой
Microsoft	0,811048571	0,42596952	0,601056319
Disney	0,889095701	0,484619629	0,288981182
Apple	0,556785876	-0,05513076	0,136985278
Amazon	0,420445519	-0,036780759	-0,04534616

15

Чтобы убедиться что существует корреляция между индексом P/E и настроениями пользователей, был произведён корреляционный анализ по Пирсону. Из таблицы мы можем увидеть. Что корреляция довольно высокая, следовательно зависимость присутствует.

Предсказательная модель

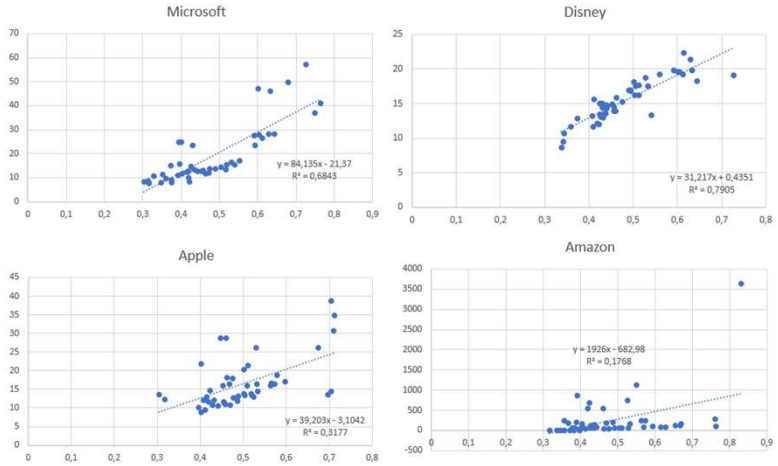


16

Чтобы строить предсказательную модель, вначале нужно убедиться, что настроения пользователей влияют на индекс P/E, а не наоборот. Для этого посмотрим на более детальный график компании Microsoft. Можно заметить, что настроения пользователей опережают индекс P/E. Графики других компаний показывают примерно такую же картину.

Линейная регрессия

Анализ:
- R^2
- диаграммы разбросов
- диаграммы остатков
- графики зависимоти остатков от времени
- диаграммы нормального распределения остатков



17

На основе полученных результатов, была построена простая предсказательная модель с помощью линейной регрессии. Был произведён анализ различных статистических метрик. Линейная регрессия показывает хорошие результаты только для отдельных компаний. Можно сделать предположения, что есть скрытые от обычных людей факторы, которые не поддаются прогнозиванию, но которые существенно влияют на показатели компании. Стоит учитывать, что в модели мнения различных людей никак не ранжируются. То есть более значимые личности как для отдельной компании, так и для мировой экономики могут оказывать значительное влияние на фондовый рынок, чем значительно большее количество обычных инвесторов и пользователей. Поэтому для некоторых компаний мы можем видеть не такие хорошие результаты.

Результаты

Собран значительный объем информации о мнениях сообщества

Показана корреляция между индексом P/E и настроениями в Twitter

Реализована модель для предсказания индекса P/E

18

Поставленные цели были выполнены. Как было замечено, предсказательная модель не показала хороших результатов, но такая цель и не стояла. Исследование показало, что мы можем применять методы анализа тональности тестов для предсказания индекса P/E. Результаты дают возможность для дальнейших исследований и построения более качественной предсказательной модели. Спасибо за внимание.