Санкт-Петербургский государственный университет Математическое обеспечение и администрирование информационных систем

Холодаева Екатерина

Sentiment Analysis using NLP to predict PE Ratio

Курсовая работа

Научный руководитель:
доц Д. А. Григорьев.

Санкт-Петербург, 2019 г.

Содержание

1 Введение **2**

2 Постановка задачи **3**

3 DataSet **4**

3.1 Используемый датасет 4

3.2 Очистка данных и предварительная обработка текста 4

4 Обзор методов **5**

5 Bag Of Words **6**

6 Word2Vec **7**

7 Bag Of Centroids **9**

8 Классификация **10**

9 Анализ результатов **11**

10 Заключение **12**

1 Введение

Классификация текстов – одна из областей обработки натуральных языков (англ. Natural Language Processing). Со временем она становится все более и более важной и перспективной. Информатизация населения и перевод текстов в электронный вид (например, электронный документооборот в Российской Федерации [**?**]) приводят к необходимости разработки эффективных алгоритмов анализа и классификации этих текстов.
Одна из областей классификации – анализ тональности текста – обработка естественного языка, классифицирующая тексты по эмоциональной окраске. Такой анализ можно рассматривать как метод количественного описания качественных данных, с присвоением оценок настроения. Противодействие терроризму, автороведческое исследование документов в криминалистике, оценка качества товаров и услуг на основе отзывов пользователей Интернет-ресурсов – в этих и во многих других областях применение анализа тональности имеет важное практическое значение.
Одна из сфер применения анализа — это анализ ситуации на фондовых рынках и прогнозирование волатильности финансовых активов.
Индекс P/E это один из самых важных показателей для инвесторов. Он показывает соотношение между ценой акции и прибылью компании и рассчитывается по формуле:

$$\mathbf{P/E} = \frac{\textbf{Share price}}{\textbf{Earnings per Share}}$$

Коэффициент используется для оценки компаний и определения того, являются ли они переоцененными или недооцененными., а также иллюстрирует срок окупаемости вложений и соразмерность прибыли. На рисунке 1 изображен график изменения индекса P/E с 2013 по 2019 год компании Facebook.

Рис. 1:

Подход к анализу P/E в этом случае не однозначный — чем выше P/E, тем менее привлекательна акция с точки зрения ее текущей доходности; с другой стороны, большое (на фоне остальных) P/E показывает, что инвесторы пророчат компании больший рост прибыли. Очень низкое же значение коэффициента может указывать на скрытые (или явные) угрозы для компании или неясность перспектив ее развития.

2 Постановка задачи

Целью данной курсовой работы является реализация методов анализа тональности текста и применение их для прогнозирования коэффициента PE. Указанная цель достигалась путем решения следующих основных задач:

1. Изучение основных методов анализа тональности текста
2. Их применение к данной задаче
3. Анализ полученных результатов

3 DataSet

3.1 Используемый датасет

В качестве обучающего датасета использовался файл, содержащий примерно 75000 предложений, имеющих положительную или отрицательную окраску., а в качестве тестового датасета набор отзывов о компании Facebook с сайта Trustpilot [?]. Это веб-сайт отзывов потребителей, основанный в Дании в 2007 году, на котором размещены обзоры компаний по всему миру. На рисунке 2 изображен пример обучающих данных.

Рис. 2:

3.2 Очистка данных и предварительная обработка текста

Для дальнейшей работы с данными необходимо провести предварительную обработку текста, а именно:

1. Удаление HTML-разметки
2. Удаление цифр и знаков препинания с помощью регулярных выражений
3. Удаление часто встречающихся, но ничего не значащих слов (стоп-слова(“the”,”a” и др))
4. Стемминг – процесс выделения основы слова для заданного слова. Стемминг используется для приведения всех однокоренных слов к единому виду, так чтобы все однокоренные слова в программе выглядели идентично.

4 Обзор методов

Анализ настроений – это область обработки естественного языка (NLP), которая строит модели, задача которых идентифицировать и классифицировать атрибуты выражения, например:
Полярность: если говорящий выражает положительное или отрицательное мнение,
Предмет: о чем идет речь,
Автор высказывания: физическое или юридическое лицо, выражающее мнение.
В мире ежедневно генерируется более миллиарда байт информации, поэтому анализ настроений — это один из ключевых инструментов для обработки этих данных. Сделав обобщение, можно разделить существующие подходы на следующие категории:

1. Подходы, основанные на правилах
2. Подходы, основанные на словарях
3. Машинное обучение с учителем
4. Машинное обучение без учителя

Суть подходов основанных на правилах состоит в том, что система делает заключение о тональности текста по набору заранее заданных правил. Например, для предложения “Я люблю рисовать” применяется следующее правило:
В предложении содержится глагол “люблю” с положительной окраской и нет отрицаний, следовательно, тональность предложения положительная.
Подходы, основанные на словарях, используют так называемые тональные словари (affective lexicons) для анализа текста. Один из примеров такого словаря— это список слов со значением тональности для каждого слова.

Алгоритм анализа текста:

1. Определяем тональность каждого слова в тексте (берем значения из словаря)
2. Вычисляем тональность текста одним из способов (один из самых простых методов — среднее арифметическое значений тональностей всех слов, или более сложный метод — обучить классификатор)

Машинное обучение с учителем является наиболее распространенным методом, используемым в исследованиях. Его суть состоит в том, чтобы обучить машинный классификатор на коллекции заранее размеченных текстах, а затем использовать полученную модель для анализа новых документов. С помощью этого метода проводится данное исследование.

Машинное обучение без учителя представляет собой наименее точный метод анализа тональности. Одним из примеров данного метода может быть автоматическая кластеризация документов.
На основании обзора существующей литературы [?, ?, ?, ?, ?] было выделено 3 основных метода, которые могут быть полезны для решения задачи: Bag Of Words, Word2Vec, Bag of Centroids.

5 Bag Of Words

Мешок слов (или Bag of Words) — это модель текстов на натуральном языке, в которой каждый документ или текст выглядит как неупорядоченный набор слов без сведений о связях между ними. Его можно представить в виде матрицы, каждая строка в которой соответствует отдельному документу или тексту, а каждый столбец — определенному слову. Ячейка на пересечении строки и столбца содержит количество вхождений слова в соответствующий документ.

Алгоритм построения модели:

1. Составляем словарь из всех слов, содержащихся в документе (стоп-слова удалены при предварительной обработке текста)
2. Затем каждому документу сопоставляем вектор, размерность которого равна мощности словаря, а каждая компонента — это количество вхождений данного слова в этот текст

Например, если даны документы:

1. I often draw in the morning. All people in the morning somewhere in a hurry.
2. I do not like to draw.

Словарь будет выглядеть следующим образом:{I, often, draw, in, the, morning, all, people, somewhere, a, hurry, do, not, like, to}

Вектор, соответствующий 1 документу — [1, 1, 1, 3, 2, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0]
Вектор, соответствующий 2 документу — [1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1] С результатами применения метода Bag Of Words можно ознакомиться в таблице 1. На рисунке 3 отображено изменение тональности отзывов пользователей о компании Facebook с 2013 по 2019 год, вычисленное с помощью метода Bag Of Words

Рис. 3:

Таблица 1:

Date	Sentiment
апр.13	1,0000
июл.13	1,0000
окт.13	0,5000
янв.14	0,4700
апр.14	0,7000
июл.14	0,4700
окт 14	0,8000
янв.15	0,8000
апр.15	0,7000
июл.15	0,7500
окт 15	0,8000
янв.16	0,9000
апр.16	0,6000
июл.16	0,9265
окт 16	0,7500
янв.17	0,8800
апр.17	0,8889
июл.17	0,9000
окт 17	0,8235
янв.18	0,8182
апр.18	0,8462
июл.18	0,8000
окт 18	0,7500
янв.19	0,8667

6 Word2Vec

Word2Vec, разработанный в 2013, году представляет собой реализацию нейронной сети, которая изучает распределенные представления для слов. Другие глубокие или рекуррентные архитектуры нейронных сетей были предложены для изучения представлений слов до этого, но главной проблемой с ними было длительное время, необходимое для обучения моделей. Word2Vec быстро учится относительно других моделей.
Для обучения Word2Vec лучше не удалять стоп-слова, поскольку алгоритм использует более широкий контекст предложения для получения высококачественных векторов слов.
Word2Vec ожидает отдельные предложения, каждое из которых представляет собой список слов. Другими словами, формат ввода представляет собой список списков.
Word2Vec фиксирует контекст слов, одновременно уменьшая размер данных.
Word2Vec – это не один алгоритм, он включает в себя две модели обучения: «Continuous Bag of Words» (CBOW) и Skip-gram. CBOW – «непрерывный мешок со словами», архитектура, которая предсказывает текущее слово, исходя из окружающего его контекста. Архитектура типа Skip-gram действует иначе: она использует текущее слово, чтобы предугадывать окружающие

его слова.4 Используемая в данной работе архитектура — скип-граммы (так как путем экс-

Рис. 4:

периментов выяснилось, что она работала медленнее, но давала более точные результаты). Метод использует искусственные нейронные сети в качестве алгоритма классификации. Первоначально каждое слово в словаре является случайным N-мерным вектором. Во время обучения алгоритм изучает оптимальный вектор для каждого слова.
Путем экспериментов выяснилось, что оптимальный размер окна - 10 слов. Размерность слов в векторе: больше функций приводит к увеличению времени выполнения и часто, но не всегда, к получению лучших моделей. Разумные значения могут быть от десятков до сотен. Здесь было использовано 300, а используемый алгоритм обучения — softmax. Алгоритм построения модели:

1. Строится словарь из наиболее часто встречающихся в документах слов (количество слов в словаре задается вручную).
2. Каждому слову в словаре сопоставляется частота его встречаемости в документах.
3. Строится дерево Хаффмана для слов
4. Субсэмплирование наиболее частотных слов (sub-sampling). Субсэмплирование — это процесс изъятия наиболее частотных слов из анализа, что ускоряет процесс обучения алгоритма и способствует значительному увеличению качества получающейся модели.
5. Для этих слов применяется алгоритм softmax (negative sampling работает быстрее, но для работы с не очень частотными словами больше подходит Hierarchical Softmax)

Одна из проблем с набором обучающих данных — это тексты переменной длины. Поскольку каждое слово является вектором в 300-мерном пространстве, можно использовать векторные операции для объединения слов в каждом тексте. Метод, который использован в этой работе – усреднение векторов слов в данном абзаце.
С результатами применения метода Word2Vec можно ознакомиться в таблице 2. На рисунке 5 отображено изменение тональности отзывов пользователей о компании Facebook с 2013 по 2019 год, вычисленное с помощью метода Bag Of Words

Рис. 5:

Таблица 2:

Date	Sentiment
апр.13	1,0000
июл.13	0,9000
окт.13	0,6000
янв.14	0,8000
апр.14	0,8000
июл.14	0,7000
окт 14	0,6667
янв.15	0,8000
апр.15	0,7000
июл.15	0,5000
окт 15	0,8000
янв.16	0,8000
апр.16	0,8000
июл.16	0,6618
окт 16	0,2500
янв.17	0,6000
апр.17	0,5556
июл.17	0,6000
окт 17	0,5882
янв.18	0,6364
апр.18	0,5385
июл.18	0,6000
окт 18	0,5417
янв.19	0,6000

7 Bag Of Centroids

Word2Vec создает кластеры семантически связанных слов, поэтому другой возможный подход заключается в использовании сходства слов в кластере. Группирование векторов таким способом называется «векторным квантованием». Для этого сначала нужно найти центры кластеров слов, что можно сделать, используя алгоритм кластеризации, такой как K-средние.
Метод проб и ошибок показал, что небольшие кластеры, в среднем всего около 5 слов или около того на кластер, дали лучшие результаты, чем крупные кластеры с большим количеством слов.
Теперь у нас есть кластерное (или «центроидное») назначение для каждого слова, и можно определить функцию для преобразования текстов в мешки центроидов. Это работает так же, как Bag of Words, но использует семантически связанные кластеры вместо отдельных слов.
С результатами применения метода Bag Of Centroids можно ознакомиться в таблице 3. На рисунке 6 отображено изменение тональности отзывов пользователей о компании Facebook с 2013 по 2019 год, вычисленное с помощью метода Bag Of Centroids

Таблица 3:

Date	Sentiment
апр.13	1,0000
июл.13	0,8000
окт.13	0,8182
янв.14	0,7000
апр.14	0,7000
июл.14	0,6000
окт 14	0,6364
янв.15	0,6667
апр.15	0,6667
июл.15	0,5385
окт 15	0,3333
янв.16	0,5455
апр.16	0,4000
июл.16	0,5147
окт 16	0,4444
янв.17	0,6000
апр.17	0,5556
июл.17	0,5455
окт 17	0,5556
янв.18	0,4545
апр.18	0,4615
июл.18	0,4200
окт 18	0,4583
янв.19	0,4667

Рис. 6:

8 Классификация

Получив векторы функций для каждого документа, впоследствии был применен классификатор Random Forest для изучения настроений.
Random Forest — алгоритм машинного обучения, заключающийся в использовании ансамбля решающих деревьев
Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе дает очень невысокое качество классификации, но за счет их большого количества результат получается хорошим.

9 Анализ результатов

Теперь сравним результаты. На основании графиков 7 и данных 4 можно сказать что в рамках задачи прогнозирования индекса P/E наилучшие результаты показал метод Bag Of Centroids, так как данная модель существенно понижает размерность вектора признаков, а также уменьшает вычислительные затраты при обучении.
При использовании модели “Мешок слов” корреляция графиков прослеживается, но довольно незначительная, и с увеличением количества данных увеличится размер корпуса, следовательно, и размерность векторов, что приведет к высокой вычислительной сложности.
Word2Vec без использования сходства слов в кластерах показал также неплохие результаты.

Рис. 7:

Таблица 4:

Method	Correlation
Bag Of Words	0,268267
Word Vector	0,507743
Bag Of Centroids	0,641367

10 Заключение

Таким образом, можно подвести итоги курсовой работы. В течение семестра были решены следующие задачи:

1. Реализованы основные методы анализа тональности текста
2. Прогнозирование индекса P/E с помощью реализованных методов
3. Произведен анализ результатов