ОЦЕНКА БИЗНЕСА С ИСПОЛЬЗОВАНИЕМ ТЕКСТОВЫХ ДАННЫХ

Коклев Пётр Сергеевич
аспирант кафедры теории кредита
и финансового менеджмента
Санкт-Петербургский
государственный университет

Аннотация: В статье рассматривается использование текстовых данных (описание компании) для присвоения справедливой рыночной капитализации предприятия. Применение простой репрезентации документа – мешка слов (bag of words) вкупе с использованием градиентного бустинга на основе деревьев решений (*GBDT*) позволяет систематически формировать качественную оценку рыночной капитализации для корпораций тестового множества. Тестовый p-квадрат для компаний *NYSE* и *NASDAQ* достигает 55.87%, а средняя абсолютная процентная ошибка прогноза всего 9.58%. Таким образом, текстовые данные обладают большим потенциалом для оценки бизнеса.

Ключевые слова: оценка бизнеса, текстовые данные, NLP, bag of words, машинное обучение, GBDT, градиентный бустинг.

BUSINESS VALUATION USING TEXT DATA

Koklev Petr Sergeevich

Abstract: The article discusses the use of textual data (company description) to assign a fair market capitalization of an enterprise. The use of a simple document representation – a bag of words model, coupled with the use of gradient boosted decision trees (GBDT) allows generate precise business valuations. The test set r-squared for the *NYSE* and *NASDAQ* companies reaches 55.87%, and the average absolute percentage error of the forecast is only 9.58%. Thus, textual data has great potential for business valuation.

Key words: business valuation, text data, NLP, bag of words, machine learning, GBDT, gradient boosting.

Введение

Текстовые данные уже довольно давно успешно используются для решения различных прикладных финансовых задач. Однако их применение для оценки бизнеса весьма ограничено. Более того, изучение существующей литературы не выявило случаев использования текстовых данных для регрессионного моделирования капитализации предприятия с целью его оценки. Настоящая работа позволит, в некоторой степени, заполнить существующий вакуум.

В первом параграфе рассматривается краткий обзор существующей литературы применения методов обработки текстовых данных (*NLP*) к проблемам корпоративных финансов. Второй параграф посвящен практическому примеру использования *NLP*, в рамках которого с применением градиентного бустинга на основе деревьев решений оценивается рыночная капитализации корпораций, котирующихся на биржах *NYSE* и *NASDAQ*. Независимой переменной, с помощью которой моделируется рыночная капитализация, выступает описание компании, предоставленное сервисом финансовой информации *Alpha Vantage*. В заключении подводятся итоги работы.

Использование текстовых данных в финансах

Успешные случаи использования текстовых данных к различным задачам корпоративных финансов позволяют предположить, что они также могут быть применены и для оценки стоимости компании. Группа отечественных авторов предложила использовать метод кодировки корпуса текста (*мешка слов*) для извлечения релевантной для стоимости акции информации из обращений генеральных директоров компаний – *CEO letters* [1]. К сожалению, крайне низкая выборка (всего 50 писем), не позволяет делать каких-либо далекоидущих выводов относительно перспектив использования такого подхода. Также существуют работы, анализирующие транскрипты телеконференций, посвящённых финансовой деятельности компаний (*earnings calls*). В работе [2], с помощью 62,664 транскриптов и методов *NLP*, авторы измеряют корпоративную культуру предприятий. З.Ке, Б. Келли и Ч. Даченг с помощью предложенного в исследовании алгоритма обработки текста *SESTM*, создают модели, идентифицирующие положительные и негативные слова из документа [3]. Каждому документу присевается числовой рейтинг, который является хорошим предиктором движения стоимости акции. Т. Логан и Б. МакДональд в пространном обзоре рассматривают многие другие последние

достижения анализа текста в области финансов [4]. Превосходный обзор методов анализа текстовых данных в экономике также предлагается в работе [5]. Таким образом, есть все основания предполагать, что текстовые данные обладают высоким потенциалом для оценки стоимости предприятия.

Пример использования текстовых данных для оценки компаний *NYSE* и *NASDAQ*

В данном разделе на конкретном примере демонстрируется потенциал использования текстовых данных для оценки бизнеса. Провайдер финансовой данных *Alpha Vantage* помимо данных финансовой отчетности компаний также предоставляет описание для каждой компании – некоторый нарратив, характеризующий корпорацию, её деятельность. Например, для компании *Tesla Inc*. описание выглядит следующим образом:

"Tesla, Inc. is an American electric vehicle and clean energy comp any based in Palo Alto, California. Tesla's current products inclu de electric cars, battery energy storage from home to grid-scale, solar panels and solar roof tiles, as well as other related produc ts and services. In 2020, Tesla had the highest sales in the plug- in and battery electric passenger car segments, capturing 16% of t he plug-in market (which includes plug-in hybrids) and 23% of the battery-electric (purely electric) market. Through its subsidiary Tesla Energy, the company develops and is a major installer of sol ar photovoltaic energy generation systems in the United States. Te sla Energy is also one of the largest global suppliers of battery energy storage systems, with 3 GWh of battery storage supplied in 2020."

Итоговый корпус документов состоит из описания для около четырех тысяч компаний. Облако наиболее часто используемых слов представлено на (рис. 1). Используя методы *NLP*, продемонстрируем, как рыночная капитализация предприятия может быть спрогнозирована исключительно на основе текстовых данных, без использования каких-либо других переменных. Корпусы документов, соответствующие каждой из шести ежеквартальных временных меток из интервала «2021/03 – 2022/06», разделяются на два несовместных множества – тренировочное и тестовое (75% и 25% соответственно). Данные тренировочного множества используется для обучения модели, а тестового – для оценки качества.

Для того, чтобы текстовые данные с описанием компаний могли использоваться в качестве независимой переменной для обучения моделей машинного обучения необходимо их трансформировать в числовое представление. Для этого будет использоваться самый распространенный и простой способ представления текста – мешок слов (*bag of words*), в рамках которого документ представляется в виде мультимножества. В результате применения мешка слов корпус текста отображается в виде так называемой терм-документной матрицы (*Document-term matrix*). Строкам матрицы соответствуют отдельные компании, а столбцам – используемые в корпусе слова. Каждой строке и столбцу соответствует значение частоты использования слова в описании компании.

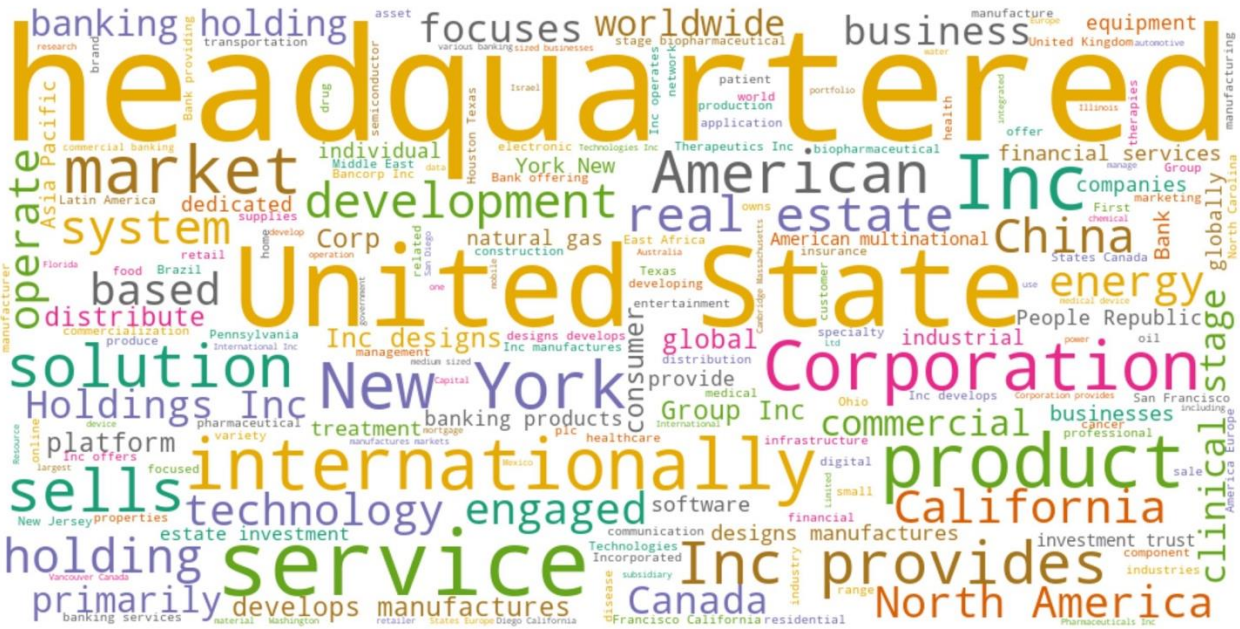

Рис. 1. Облако наиболее часто используемых слов в корпусе документов. Размер слова отражает частоту его использования

Такая репрезентация позволяет сохранить информацию о словах и об их количестве в документе, однако информация о последовательности слов теряется. Кроме этого, предлагается использовать следующие процедуры предварительной обработки текстовых данных, позволяющие извлечь больше

информации из текста и улучшить точность прогнозов по сравнению с применением мешка слов в базовом виде.

1. Токенизация документов с применением лемматизации. Лемматизация – процесс приведения различных форм одного слова к единому представлению для того, чтобы они могли быть анализированы как единое целое. К примеру, словам «studies», «studying» будет присвоена единая лемма – «study». Так, модель будет трактовать слова «studies» и «studying» как один токен – «study». По своей сути лемматизация является близким аналогом процесса определения корня слова в русском языке. Естественно, лемматизация всех слов из корпуса вручную невозможна. Для этого используется библиотека *NLTK*. В итоге, столбцам терм-документной матрицы на самом деле соответствуют не слова, а токены.

2. Исключение стоп-слов (шумовых слов). Стоп-слова (*stop words*) – список слов, исключаемых при обработки текстовых данных. Как правило, список состоит из наиболее часто употребляемых в языке слов, которые можно встретить практически в каждом документе. Следовательно, эти слова не имеют ценности, так как бесполезны при дифференциации компаний. Список шумовых слов также представлен библиотекой *NLTK*.

3. Использование n-грамм. Эта техника позволяет частично восстановить информацию, заложенную в порядке слов документа. После применения частного случая n-грамм – биграммы (при $n = 2$), столбцы терм-документной матрицы представляют не только отдельные токены, но и последовательности из двух токенов, встречающиеся в корпусе.

Результаты оценки точности прогнозов капитализации для компаний тестового множества с использованием метода машинного обучения – градиентного бустинга на основе деревьев решений (*GBDT*) показаны в (Таблица). Несмотря на то, что точность оценки сильно зависит от случайного разбиения данных на тестовое и обучающее множества, для каждой из шести временных меток, использование описания компании значительно превосходит наивный прогноз капитализаций (положительный p-квадрат для каждого момента времени). Разброс средней абсолютной ошибки прогноза варьируется от 9.58% для последней временной отметки и 19.48% для марта 2021 года.

Таблица 1

Качество моделей прогнозирования рыночной капитализации с использованием текстовых данных

Дата	Число компаний обучающего множества	Число компаний тестового множества	$MAPE_{oos}$	R^2_{oos}
2021-03	2815	939	19.48%	9.41%
2021-06	2991	998	14.89%	1.76%
2021-09	3024	1009	11.39%	55.87%
2021-12	3051	1018	11.99%	35.94%
2022-03	2934	978	11.82%	42.49%
2022-06	2901	967	9.58%	48.73%

В работе [6], медианная ошибка прогноза стоимости компаний, данная студентами-старшекурсниками финансового факультета, составила 41.3%. Это значительно выше, чем средняя ошибка прогноза для полученной модели даже для самого «неудачного» квартала. Такие результаты особенно знаменательны с учетом того, что, как очевидно, оценщики в реальном мире не ограничены данными описания компаний. Более того, короткое описание компании, представленное сервисом финансовой информации, скорее всего никак не будет использовано оценщиками в рамках традиционных *DCF* моделей.

Конечно, полученные результаты следует воспринимать с известной долей скептицизма. В отличие от использования финансовой отчетности, можно привести аргументы в пользу того, что то или иное текстовое описание компании является следствием капитализации компании, а не наоборот. Это не является серьёзной проблемой, так как целью данной работы является демонстрирование потенциала использования текстовых данных для оценки стоимости компании, а не строгое статистическое тестирование какой-либо выдвигаемой научной гипотезы (помимо гипотезы о высоком потенциале использования текстовых данных). Кроме того, так или иначе, высокая точность прогнозов получена именно для компаний отдельного тестового множества. То есть компаний, текстовое описание которых никак не использовалась при обучении модели.

Результаты применения методов оценки важности признаков также косвенно указывают на то, что использование описания компаний может

успешно применяться в текущем виде. Некоторые ключевые для точности прогнозов токены лишены каких-то оценочных суждений, субъективности. То, что компания американская (токен «american») или является бывшей дочерней компанией (токен «former subsidiary») является неокрашенными, сухими фактами. Результаты относительной важности токенов, полученные двумя различными методами оценки важности признаков (*prediction value change* и *SHAP*) представлены на (рис. 2) и (рис. 3). Заметим, что списки наиболее ценных для прогнозов токенов, определенные разными методами оценки важности признаков во многом пересекаются. Например, токены «multinational», «american», «world valuable», «world», «big five» можно встретить на обеих диаграммах.

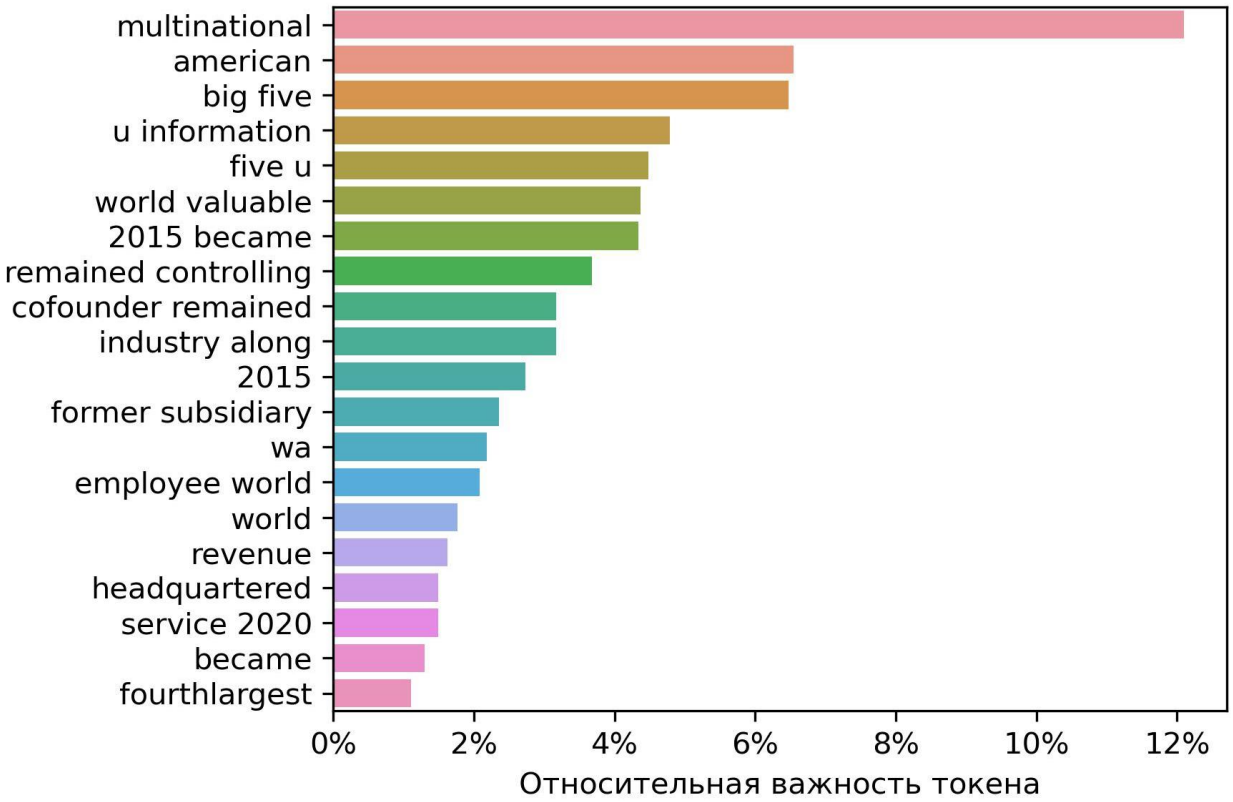

Рис. 2. Наиболее важные токены для оценки стоимости компании. Метод Prediction Value Change

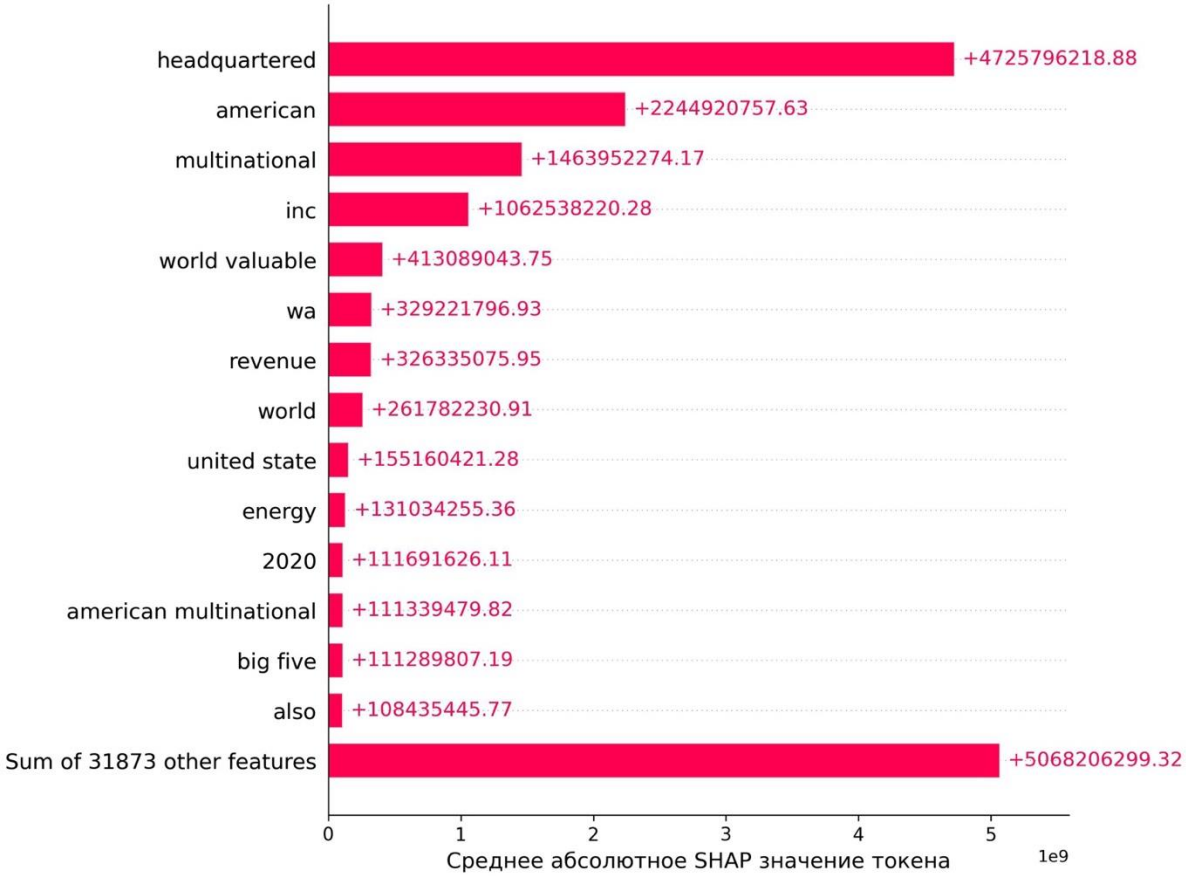

Рис. 3. Важность токенов, оцененная методом *SHAP*

Наконец, с целью иллюстрации важного принципа, сформируем оценку рыночной стоимости собственного капитала для российской компании *Beluga Group*. Описание компании на официальном сайте выглядит следующим образом:

"BELUGA GROUP — крупнейшая алкогольная компания в России, лидер по производству водки и ликеро-водочных изделий на этом рынке, а также один из главных импортеров крепкого алкоголя в стране. В основе философии нашего бизнеса лежат ориентация на потребителя, лидерство, инновации, глобальность и возможность самореализации. Это — наша «ДНК», благодаря которой мы всегда движемся вперед."

После перевода на английский язык с использованием сервиса *Google Translate*, описание компании может использоваться для прогнозирования. Итоговый прогноз капитализации компании составил более двадцати миллиардов долларов, что значительно больше фактической капитализации.

Такая чудовищная ошибка прогноза иллюстрирует важный принцип применения методов машинного обучения. Обучающие данные и данные для прогнозирования должны принадлежать одному и тому же процессу. Предвзятое описание компании, полученное с официального сайта компании не может быть использовано. Так, для применения обученной модели, описание компании *Beluga Group* должно быть получено из того же распределения, что и описание компаний обучающей выборки. Разумнее всего получить описание компании для оценки от поставщика финансовой информации *Alpha Vantage*.

Заключение

На примере корпораций *NYSE* и *NASDAQ* удалось показать, что использование текстовых данных обладает огромным потенциалом для оценки стоимости компании. Не используя данных финансовой отчетности, макроэкономических переменных, современные методы обработки текстовых данных и машинного обучения позволяют объяснить около половины дисперсии рыночной капитализации предприятий. Средняя абсолютная ошибка прогноза, в свою очередь, не превышает 20%. Таким образом, расширение набора независимых переменных текстовыми данными позволит поднять точность регрессионных моделей оценки бизнеса на качественно иной уровень.

Список литературы

1. Е. А. Федорова, И. С. Демин, Л. Е. Хрустова, Р. А. Осетров и Ф. Ю. Федоров, «Влияние тональности писем CEO на финансовые показатели компании,» *Российский журнал менеджмента,* т. 15, № 4, pp. 441-462, 2017.

2. K. Li, F. Mai, R. Shen и X. Yan, «Measuring Corporate Culture Using Machine Learning,» *The Review of Financial Studies,* т. 34, № 7, p. 3265–3315, 2021.

3. Z. T. Ke, B. T. Kelly и D. Xiu, «Predicting returns with text data,» *National Bureau of Economic Research,* 2019.

4. T. Loughran и B. McDonald, «Textual Analysis in Accounting and Finance: A Survey,» *Journal of Accounting Research,* т. 54, № 4, pp. 1187-1230, 2016.

5. M. Gentzkow, B. Kelly и M. Taddy, «Text as Data,» *Journal of Economic Literature,* т. 57, № 3, pp. 535-74., 2019.

6. P. Geertsema и H. Lu, «Machine Valuation,» *Available at SSRN: https://ssrn.com/abstract=3447683,* 2019.