ОПТИМАЛЬНЫЕ МЕТОДЫ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ НОВОСТНОГО ПОТОКА ДЛЯ ОБНАРУЖЕНИЯ СТАТЕЙ, ВЛИЯЮЩИХ НА КРЕДИТНЫЕ РИСКИ

М. Ю. Беляев

Компания «Deloitte», Москва, Россия
E-mail: mbelyaev@deloitte.ru

Целью исследования является поиск оптимального алгоритма фильтрации новостного потока, который может иметь влияние на эффективность работы моделей оценки вероятности дефолта компаний российского рынка, применяемых финансовыми институтами. Научная новизна исследования заключается в разработке первого этапа построения конвейера новостных статей, который позволит в автономном, не требующем внимания человека онлайн режиме оценивать возможность компании платить по своим обязательствам. В результате был найден алгоритм, позволяющий с достаточно быстрой скоростью присваивать каждой статье определенную заранее тему на основании содержащихся в ней слов.

OPTIMAL METHODS OF NEWS FLOW TOPIC MODELLING FOR INDICATING ARTICLES, WHICH INFLUENCE ON CREDIT RISKS

M. Yu. Belyaev

The objective of this research is to find an optimal filtering algorithm for the news stream, which may have an impact on the efficiency of the models used by financial institutions for assessing probability of Russian companies default. The scientific novelty of the research lies in the building first stage of news articles conveyor, which will allow assessing companies' ability to pay on its obligations in an autonomous online mode that does not require human attention. As a result, there was found algorithm that allows assigning a predetermined topic to each article based on the words it contains.

Введение.

Постоянно возрастающий уровень конкуренции требует от финансовых институтов увеличение скорости принятия решений относительно благонадежность компаний-контрагентов. Внешние оценки рейтинговых агентств покрывают лишь незначительную долю российского рынка, а также сформированы с учетом анализа данных за относительно длительный промежуток времени (как правило, три года) и не отражают вероятность дефолта контрагента в текущий момент времени. Финансовые институты вынуждены инвестировать в разработку собственных моделей оценки компаний. При этом корректировка финального рейтинга до сих пор остается привилегией риск-аналитика. При принятии решения о корректировки сотрудник анализирует новостной фон контрагента.

Умение встраивать логику анализа риск-аналитиков в систему автоматического принятия решения, равно как и выделение факторов, оказывающих су-

щественное влияние при оценке вероятности дефолта, является до сих пор нерешенной задачей, но весьма актуальной задачей, решение которой позволило бы:

1) Элиминировать человеческий фактор при принятии решения о корректировке рейтинга;

2) Снизить издержки за счет автоматизации корректировки.

Для решения данных задач, был собран корпус из 126 176 статей по 2920 компаниям, затем над данным корпусом была произведена предобработка и векторизация слов по методу “bag-of-words” и были построены две модели: модель Латентного Распределения Дирихле (LDA) и модель с аддитивной регуляризацией (ARTM).

Предобработка текста.

Предобработка корпуса проводилась в несколько этапов:

1. Очистка от наиболее часто встречающихся слов в русской речи, таких как местоимения, служебные части речи, предлоги, междометия и т.д.

2. Стемминг оставшихся слов.

3. Составление используемого словаря из 10000 наиболее встречающихся слов, длина которых не меньше 3 символов.

4. Векторизация статей с размерностью используемого словаря словарю, где $n_{i,j}$ – количество появлений j-го слова в документе i

Модель LDA.

LDA – трехуровневая иерархическая Байесовская модель, в которой каждая статья представлена как сочетание заданного набора тем. Каждая тема, в свою очередь, представлена как сочетание элементов заданного словаря. [1]

Под темой/топиком в данной работе понимается набор терминов, то есть слов или словосочетаний, которые совместно часто встречаются в документах. Каждая тема характеризуется словарем и своими вероятностями терминов из этого словаря, p(w|t) — вероятность термина w в теме t.

Под тематикой документа мы понимаем условное распределение p(t|d), где t — тема, а d — документ.

Предпосылки модели [2]:

1. Порядок новостей не важен

2. Порядок терминов в статьях не имеет значения (гипотеза «мешка слов»)

3. Каждое слово статьи связано с некоторой темой, то есть каждая пара (d, w) связана с некоторой темой $t \in T$. Следовательно, коллекция документов представляет собой последовательность троек (d, w, t), в которой темы являются латентными: они не видны и для их определения как раз используется тематическая модель.

4. Гипотеза условной независимости: p(w|t,d) = p(w|t) заключается в том, что вероятность слова в статье определяется только темой, а не самой статьей.

5. документ относится к небольшому числу тем.

6. каждая тема состоит из небольшого числа терминов, лексического ядра, которое существенно отличает эту тему от остальных.

7. Матрицы вероятностей топиков относительно документов и слов относительно топиков порождаются распределениями Дирихле.

Оптимальные параметры для данной модели находились с помощью кросс-валидации и поиска комбинации заранее заданных параметров, которые при лучшем сочетании давали бы минимальную перплексию.

Модель ARTM.

В качестве оппонента LDA была взята модель ARTM [2], которая основывается предпосылках 1-6 модели LDA (модель PLSA [3]), но к максимизируемой функции логарифма правдоподобия L(Φ, Θ) из [2] прибавили регуляризатор разреженности матрицы Φ с коэффициентом τ1=-0.1 и регуляризатор декоррелирования тем в матрице Φ с коэффициентом τ2=1.5e+5

Построение моделей и валидация.

В качестве языка программирования был выбран Python, поскольку он обладает широким спектром разработанных библиотек для разработки моделей машинного обучения на текстовых данных. Для создания тематических моделей применялась библиотека bigARTM (для ARTM) и sklearn (для LDA).

Оценка перплексии после поиска наилучших параметров для Латентного Распределения Дирихле получилась равной 2357.88, а оценка той же самой метрики на модели аддитивной регуляризации с двумя регуляризаторами без поиска оптимальных параметров была равна 2109.81

Все параметры, которые присутствуют в описаниях [4, 5] к классам sklearn.decomposition.LatentDirichletAllocation и artm.ARTM, но которые отсутствуют в перечне ниже определятся по умолчанию документацией:

- Параметры расчета LDA:

1. n_components = 15
2. doc_topic_prior = 1/10
3. learning_decay = 0.7

- Параметры расчета ARTM:

1. num_topics = 20
2. регуляризаторы, упомянутые в разделе модель ARTM.

Заключение.

На основании приведенных исследований можно утверждать, что модель ARTM имеет больше потенциала в диверсификации тем, чем LDA, это обусловлено тем, что концепция аддитивной регуляризации дает больше возможностей по настройке модели за счет рассмотрения более общей модели PLSA. Данные открытия могут послужить начальным этапом построения data pipeline, в котором новостной поток будет преобразовываться в количественную оценку потенциала компании и сможет служить одним из значимых предикторов в оценке кредитных рисков.

СПИСОК ЛИТЕРАТУРЫ

1. *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003 P. 993–1022.

2. *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. 2014. С. 1-10.

3. *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1999. С. 50–57.

4. Документация BigARTM библиотека с открытым кодом для тематического моделирования больших коллекций текстовых документов и массивов транзакционных данных. [Электронный ресурс]. URL: https://bigartm.readthedocs.io/en/v0.8.1/index.html (дата обращения: 16.10.2020).

5. Документация sklearn [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html (дата обращения: 10.10.2020).