Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Шаповалов Никита Анатольевич

# Интерпретируемые тематические модели новостных потоков для прогнозирования на финансовых рынках

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

**Научный руководитель:**

д.ф.-м.н., доцент

К. В. Воронцов

Москва, 2018

# Содержание

**Аннотация**

В настоящее время машинное обучение активно используется для анализа финансовых рынков и в частности для написания торговых стратегий. Основной анализ разделяется на технический и фундаментальный. При техническом анализе акцент делается на поиск закономерностей в исследуемых сигналах для непосредственного предсказания сигнала в следующий момент времени.

При фундаментальном анализе основной целью является поиск закономерностей между финансовыми сигналами и разнородной информацией – от финансового состояния компаний, связанных с сигналом, до политического положения в странах третьего мира. Одним из источников информации являются потоки новостей, которые могут непосредственно влиять на цены финансовых инструментов.

Известно, что вероятностные тематические модели хорошо моделируют структуру текстов на естественном языке, выявляя слова, близкие по смыслу в конкретной предметной области. В данной работе исследуется зависимость финансовых котировок от новостей, получаемых в реальном времени, точнее, от новостных заголовков. Производится сравнение вероятностных тематических моделей для задачи предсказания движения рынка.

# 1 Введение

При анализе финансовых рынков все исследования базируются на одной из двух гипотез о поведении рынка. Модели, основанные на гипотезе эффективного рынка [5], предполагают, что вся существенная информация влияет на рынок в момент её появления и в полном объеме. При этом различают три формы эффективности – сильную, среднюю и слабую [6] – которые отличаются по степени влияния информации на рынок и времени отклика.

В свою очередь, гипотеза адаптивного рынка [9] предполагает, что на движение рынка влияет не только информация, но также и поведение участников рынка, которое также зависит от получаемой информации. Таким образом, появляется возможность использовать имеющуюся неэффективность при анализе движения рынка.

Одним из источников информации, влияющей на рынок, являются новости – от коротких записей в социальных сетях до многостраничных отчетов финансовых и аналитических ведомств [11]. Среди методов анализа зависимости рынка от новостных потоков отдельный интерес представляет тематическое моделирование [8]. Методы тематического моделирования успешно используются для предсказания цен финансовых инструментов [10] [3], для улучшения торговых стратегий [10], для выявления взаимосвязей между инструментами [4] [7]. Отдельные исследования демонстрируют успешное использование информации из социальных сетей для анализа рынка [3, 12].

В данной работе приводится исследование зависимости котировок финансовых инструментов от потока новостных заголовков с использованием методов тематического моделирования. Решается задача бинарной классификации – определения моментов времени, когда цены финансовых инструментов существенного отклоняются от ожидаемого значения.

Работа структурирована следующим образом. Раздел 2 посвящен описанию вероятностных тематических моделей – базовой постановки задачи, теории ARTM, а также набора моделей для работы с короткими текстами. Раздел 3 содержит постановку задач бинарной классификации с обзором используемых моделей. Раздел 4 описывает эксперименты, позволяющие сравнить модели бинарной классификации на двух датасетах с новостями. Раздел 5 раскрывает полученные результаты.

# 2 Вероятностные тематические модели

## 2.1 Аддитивная регуляризация тематических моделей

Пусть $D$ — *коллекция* текстовых документов, $W$ — множество употребляемых в них слов (*словарь*).

Любой документ $d \in D$ представляет собой последовательность слов из словаря:

$(w_1, \ldots, w_{n_d}), w_i \in W$. Предполагается гипотеза *мешка слов* (порядок элементов в документе не важен), благодаря которой документ отождествляется с набором $\{n_{dw} \mid w \in W\}$, представляющим собой число вхождений каждого токена $w$ в документ $d$. Также предполагается, что существует конечное множество тем $T$, и каждое появление токена $w$ в документе $d$ связано с темой $t \in T$, которая заранее не известна. Вкупе с этими предположениями, коллекция документов рассматривается как набор независимых троек $(w_i, d_i, t_i), i = 1, \ldots, n$, полученных из дискретного распределения $p(w, d, t)$ на вероятностном пространстве $D \times W \times T$.

Запишем вероятностную тематическую модель (BTM) [13, 14]:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, d \in D, w \in W. \tag{1}$$

Параметры $\theta_{td} = p(t|d)$ и $\phi_{wt} = p(w|t)$ образуют матрицы $\Theta = (\theta_{td})_{T \times D}$ – дискретные распределения тем для документов, и $\Phi = (\phi_{wt})_{W \times T}$ – дискретные распределения слов для тем. Из этого следует, что эти матрицы являются *стохастическими* (каждый столбец представляет собой дискретное распределение).

Чтобы найти неизвестные элементы матриц $\Phi, \Theta$ по наблюдаемой коллекции документов, максимизируем логарифм правдоподобия наблюдаемой коллекции:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(w|d) \to \max_{\Phi, \Theta}$$

С учетом выражения для $p(w|d)$ и стохастичности матриц получаем задачу условной оптимизации:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta}; \tag{2}$$

$$\sum_{w \in W} \phi_{wt} = 1, \phi_{wt} \geq 0; \tag{3}$$

$$\sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0, \tag{4}$$

В данном виде задача не является *корректно поставленной по Адамару*, так как в общем случае множество её решений является бесконечным (например, можно производить согласованные перестановки столбцов матрицы $\Phi$ и строк матрицы $\Theta$). Для устранения недоопределенности используют подход, который называется *регуляризацией*. Для этого добавляют дополнительный критерий – регуляризатор, который учитывает особенности задачи и знания в предметной области.

В рамках аддитивной регуляризации тематических моделей (ARTM) [13, 14] дополнительные регуляризаторы рассматриваются в виде линейной комбинации, добавляемой к основному функционалу:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \tag{5}$$

где $R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$.

Стоит отметить, что в частном случае, когда $R(\Phi, \Theta) \equiv 0$, мы получаем модель *вероятностного латентного семантического анализа* (probabilistic latent semantic analysis, PLSA) – первая вероятностная тематическая модель, предложенная Томасом Хоффманном в 1999 году [8].

**Регуляризаторы сглаживания/разреживания тем** Общий вид регуляризаторов сглаживания и разреживания [14]:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \tag{6}$$

Если $\beta_0 > 0$, то первая часть регуляризатора, соответствующая матрице $\Phi$, представляет собой сумму кросс-энтропий между темами $\boldsymbol{\phi}_t$ и фиксированным распределением $\boldsymbol{\beta} = (\beta_w, w \in W)$. Аналогично, при $\alpha_0 > 0$, $\Theta$-часть регуляризатора является суммой кросс-энтропий между темами $\boldsymbol{\theta}_d$ и фиксированным распределением $\boldsymbol{\alpha} = (\alpha_t, t \in T)$. В силу свойств кросс-энтропии максимизация данного функционала приводит к тому, что распределения $\boldsymbol{\phi}_t$ и $\boldsymbol{\theta}_d$ становятся похожи на фиксированные распределения $\boldsymbol{\beta}$ и $\boldsymbol{\alpha}$ соответственно. При выборе равномерных распределений это приводит к сглаживанию тем; в этом случае модель эквивалентна модели Латентного размещения Дирихле LDA [2].

Теперь, если $\beta_0 < 0$ и $\alpha_0 < 0$, то максимизация регуляризатора соответствует тому, что темы будут максимально удаляться от целевых распределений. При использовании равномерных распределений это приводит к тому, что темы $\boldsymbol{\phi}_t$ и $\boldsymbol{\theta}_d$ становятся разреженными.

**Регуляризатор декоррелирования** Регуляризатор декоррелирования [14] используется для уменьшения похожести (корреляции) тем между собой. Для этого минимизируется сумма попарных ковариаций между всеми парами тем:

$$R(\Phi) = -\sum_{t,s \in T} \sum_{w \in W} \phi_{wt} \phi_{ws} \tag{7}$$

**Обучение** Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Известно [13, 14], что в таком случае локальный максимум задачи (2)-(4) удовлетворяет следующей системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right); \tag{8}$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \tag{9}$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}; \tag{10}$$

где оператор $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$.

Решение системы уравнений (8)-(10) с помощью метода простой итерации эквивалентно EM-алгоритму, где на E-шаге мы пересчитываем значения $p_{tdw}$ согласно (8), а на M-шаге вычисляем $\phi_{wt}$ и $\theta_{td}$ на основе (9)-(10).

## 2.2 Мультимодальные тематические модели

В рамках *мультимодальных тематических моделей* документ, помимо слов, может содержать метаданные. Для описания документов с разнородной информацией вводится понятие *модальности* – тип метаданных со своим конечным словарем. Модальности могут быть как текстовыми (слова, словосочетания, теги, именованные сущности), так и нетекстовыми (авторы, моменты времени, классы или категории, изображения и многие другие).

В рамках теории ARTM любой документ $d \in D$ представляет собой последовательность токенов различных модальностей: $(w_1, \ldots, w_{n_d}), w_i \in W$, где $W = W^1 \sqcup \ldots \sqcup W^m$.

Тематическая модель для каждой модальности строится по аналогии с 1:

$$p(w|d) = \sum_{t\in T} p_m(w|t)p(t|d) = \sum_{t\in T}\phi^m_{wt}\theta_{td}, d \in D, w \in W^m, m = 1, \ldots, M. \tag{11}$$

Параметры $\theta_{td} = p(t|d)$ и $\phi^m_{wt} = p_m(w|t)$ образуют матрицы $\Theta = (\theta_{td})_{T\times D}$ – дискретные распределения тем для документов, и $\Phi^m = (\phi_{wt})_{W^m\times T}$ – дискретные распределения токенов каждой модальности для тем. Обозначим блочную матрицу-столбец $(\Phi^1, \ldots \Phi^m)^T$ за $\Phi$.

При построении мультимодальных регуляризованных тематических моделей решается задача максимизации взвешенной суммы логарифмов правдоподобия для каждой модальности и регуляризаторов:

$$\sum_{m=1}^{M}\tau_m\mathcal{L}_m(\Phi^m, \Theta) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}; \tag{12}$$

$$\sum_{w\in W^m}\phi_{wt} = 1, \phi_{wt} \geq 0; \tag{13}$$

$$\sum_{t\in T}\theta_{td} = 1, \theta_{td} \geq 0, \tag{14}$$

Веса модальностей $\tau_m$ позволяют учесть модальности с учетом их важности в задаче.

**Обучение** В рамках нескольких модальностей процедура обучения обобщается следующим образом. Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Известно [13, 14], что в таком случае локальный максимум задачи (12)-(14) удовлетворяет следующей системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$p_{tdw} = \underset{t \in T}{\text{norm}} \left( \phi_{wt} \theta_{td} \right); \tag{15}$$

$$\phi_{wt} = \underset{w \in W^{m(w)}}{\text{norm}} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \qquad n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw}; \tag{16}$$

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \qquad n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw}; \tag{17}$$

где оператор $\underset{t \in T}{\text{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$; $m(w)$ — модальность токена $w$: $w \in W^{m(w)}$.

## 2.3 Проблема коротких текстов

Традиционные вероятностные тематические модели, такие как PLSA, LDA и их обобщение в рамках ARTM, при обучении неявно опираются на совстречаемость слов в рамках одного документа – эти совстречаемости аккумулируются в счетчках, на основе которых пересчитываются матрицы $\Phi$, $\Theta$. В большинстве текстовых коллекций документы имеют достаточную длину для того, чтобы можно было пользоваться частотными оценками. В то же время известно, что традиционные модели плохо работают для коллекций с короткими текстами [15], когда актуальной информации мало для грамотного пересчета счетчиков (en. *severe data sparsity*). Примером коротких текстов являются сообщения социальных сетей.

В рамках решения данной проблемы интерес представляют следующие способы:

**Агрегация коротких текстов** Данный способ предполагает, что тексты в коллекции обладают дополнительными метаданными, на основе которых их можно агрегировать. Например, если документы имеют временные метки, то для агрегации можно использовать временные окна, внутри которых объединять тексты в один. В случае, если у текста есть автор или авторы, то можно группировать тексты на основе принадлежности одному и тому же автору. Если же тексты имеют ссылки друг на друга (например, в социальных сетях сообщения могут быть написаны в качестве ответа на исходное сообщение), то можно агрегировать сообщения на основе цепочек в этих графах ссылок.

**Biterm Topic Model** В модели Biterm Topic Model (BTM, [15]) совстречаемость слов в одном документе учитывается явно. Для этого используется *битерм* – пара слов, которые находятся в одном документе на небольшом расстоянии друг от друга. Обозначение: $b = (w_1, w_2) \in B$, где $B$ – множество всех битермов.

Основные предположения, на основе которых строится модель:

- битермы появляются независимо друг от друга;
- вероятность появления битерма не зависит от документа, а только от структуры самого битерма;
- как и в теории ARTM, в модели BTM битерм связан с темой $t \in T$, которую мы не наблюдаем;
- при фиксированной теме вероятность битерма равна произведению вероятностей составляющих его слов: $p(b|t) = p(w_1|t)p(w_2|t)$

Используя формулу полной вероятности, можно расписать вероятность появления битерма $b \in B$:

$$p(b) = \sum_{t \in T} \pi_t \phi_{w_1 t} \phi_{w_2 t}, \tag{18}$$

где $\pi_t$ – априорные вероятности появления тем, а $\phi_{w_i t} = p(w_i|t)$ – вероятности слов в темах.

Для построения модели решается задача максимизации логарифма правдоподобия:

$$\sum_{(w_1,w_2) \in B} \ln \sum_{t \in T} \pi_t \phi_{w_1 t} \phi_{w_2 t} \to \max_{\pi, \Phi}; \tag{19}$$

$$\sum_{w \in W} \phi_{wt} = 1, \phi_{wt} \geq 0; \tag{20}$$

$$\sum_{t \in T} \pi_t = 1, \pi_t \geq 0 \tag{21}$$

Для обучения модели авторы предлагают использовать сэмплирование Гиббса.

При сравнении на коллекциях коротких текстов модель показывает себя лучше LDA как по набору критериев качества, так и при ручной оценке получаемых топиков. В то же время, на коллекциях с документами умеренной длины BTM ведёт себя наравне с традиционным LDA.

**Word Network Topic Model** Модель Word Network Topic Model (WNTM, [16]), так же, как и BTM, учитывает совстречаемость слов, но немного по-другому. Если в предыдущей модели за основу брались слова, которые составляли битерм, то в WNTM для каждого слова рассматривают его *контекст* – набор слов, которые встречаются рядом

с этим словом в коллекции. При этом каждый контекст рассматривается как отдельный псевдо-документ. Это позволяет запускать традиционные тематические модели (PLSA/LDA/ARTM) на построенной коллекции псевдо-документов.

Формально говоря, мы получаем следующую задачу оптимизации:

$$\sum_{c\in W}\sum_{w\in W} n_{cw} \ln \sum_{t\in T} \phi_{wt}\theta_{tc} \to \max_{\Phi,\Theta}; \tag{22}$$

$$\sum_{w\in W} \phi_{wt} = 1, \phi_{wt} \geq 0; \tag{23}$$

$$\sum_{t\in T} \theta_{tc} = 1, \theta_{tc} \geq 0, \tag{24}$$

где $c \in W$ обозначает контекст слова $c$.

# 3 Задача бинарной классификации

Пусть $X$ – множество описаний объектов, $Y$ – множество меток классов. Предполагается, что существует неизвестная зависимость – отображение $y : X \to Y$, значения которой известны только для объектов обучающей выборки $X^N = \{(x_1, y_1) \dots (x_N, y_N)\}$. В задаче *классификации* требуется по имеющейся обучающей выборке построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$.

*Бинарная классификация* является частным случаем, когда множество $Y$ состоит из двух элементов. Без ограничения общности можно считать, что $Y = \{0, 1\}$.

## 3.1 Тематические модели классификации

Тематические модели классификации являются частным случаем мультимодальных тематических моделей, в которых каждый документ может содержать метки классов, относящиеся к отдельной модальности $C$. Для удобства часть матрицы $\Phi$, соответствующую модальности классов $C$ обозначают за $\Psi$.

Одно из применений тематических моделей классификации – восстановление меток классов по содержанию документа. Пусть у нас имеется коллекция документов $D_{test}$, в которой каждый документ содержит только слова $w \in W$. Тематическая модель классификации позволяет на основе этой информации получить $p(c|d), c \in C, d \in D_{test}$:

- Неизвестные $\theta_{td}$ находятся путем решения задачи условной оптимизации (12)-(14) при фиксированной матрице $\Phi$;
- Имеющиеся матрицы $\Psi$, $\Theta$ используются для вычисления условных вероятностей $p(c|d) = \sum_{t\in T} \psi_{ct}\theta_{td}$.

## 3.2 Тематические модели совстречаемости слов

Модели BTM и WNTM, которые учитывают совстречаемость слов, не позволяют учесть метки классов напрямую, поскольку они оперируют не непосредственно документами, а статистикой совстречаемости слов, посчитанной по всей коллекции.

Есть два способа решения проблемы с учетом классов:

- Обучить модели, на их основе построить тематические профили документов $\{p(t|d)\}_{t\in T}$, которые затем использовать для бинарной классификации;
- Добавить метки класса как модальности, после чего использовать полученные профили меток классов для вычисления вероятностей принадлежности документов

**Biterm Topic Model, [15]** Рассчитаем апостериорную вероятность темы $t \in T$ для документа коллекции $d \in D$:

$$p(t|d) = \sum_{b\in B} p(t|b)p(b|d) \tag{25}$$

$$p(b|d) = \frac{N_{db}}{N_d} \tag{26}$$

$$p(t|b) \propto p(t)p(b|t) = p(t)p(w_1|t)p(w_2|t) = \pi_t\phi_{w_1t}\phi_{w_2t} \tag{27}$$

**Word Network Topic Model, [16]** Аналогичным способом можно рассчитать апостериорную вероятность темы для WNTM:

$$p(t|d) = \sum_{w\in d} p(t|w)p(w|d) = \sum_{w\in d} \theta_{tw}p(w|d) \tag{28}$$

$$p(w|d) = \frac{N_{dw}}{N_d} \tag{29}$$

## 3.3 Оценивание качества бинарной классификации

При решении задач бинарной классификации основными критериями качества алгоритма $a(x)$ являются:

- Точность (*Precision*) $\dfrac{\sum\limits_{k=1}^{N} [y_k = 1]\,[a(x_k) = 1]}{\sum\limits_{k=1}^{N} [a(x_k) = 1]}$;

- Чувствительность (*Recall*) $\dfrac{\sum\limits_{k=1}^{N}[y_k=1]\,[a(x_k)=1]}{\sum\limits_{k=1}^{N}[y_k=1]}$;

- Специфичность: $\dfrac{\sum\limits_{k=1}^{N}[y_k=0]\,[a(x_k)=0]}{\sum\limits_{k=1}^{N}[y_k=0]}$;

- Площадь под ROC-кривой (*Area Under ROC, AUROC*). ROC-кривая – множество достигаемых пар (1-специфичность, чувствительность) при всевозможных значениях порога классификации.

- Площадь под кривой Precision-Recall (*AUC-PR*). кривая Precision-Recall – множество достигаемых пар (точность, чувствительность) при всевозможных значениях порога классификации.

При исследовании моделей будем использовать интегральные критерии AUROC и AUC-PR, поскольку они не зависят от порога классификации.

# 4 Эксперименты

Построение тематических моделей осуществлялось с использованием библиотеки тематического моделирования BigARTM[1] [13].

## 4.1 Предобработка исходных данных

Имеется набор новостных заголовков с июля по декабрь 2016 года включительно. Для каждого заголовка известно время опубликования соответствующей ему новости.

Новостные заголовки используются для анализа временных сигналов за аналогичный промежуток времени. Каждый временной сигнал является последовательностью цен сделок для соответствующего биржевого инструмента. Цены по сделкам агрегированы по секундам путем усреднения по объему. Стоит отметить, что сделки по инструментам совершаются не круглосуточно, а также что интервалы, в которые не торгуются инструменты, различаются от инструмента к инструменту.

Всего рассматриваются данные по 8 фьючерсным инструментам:

- **AUDUSD** – фьючерс на валютную пару *австралийский доллар/доллар США*;

- **Dax** – фьючерс на индекс *DAX*;

- **EURUSD** – фьючерс на валютную пару *евро/доллар США*;

[1] `http://bigartm.org`

- **EuroStoxx** – фьючерс на индекс *EuroStoxx50*;
- **GBPUSD** – фьючерс на валютную пару *британский фунт стерлингов/доллар США*;
- **MXNUSD** – фьючерс на валютную пару *мексиканское песо/доллар США*;
- **NZDUSD** – фьючерс на валютную пару *новозеландский доллар/доллар США*;
- **OilBrentIce** – фьючерс на нефть марки *Brent*;

В данном исследовании основной интерес представляют не сами цены, а их изменение за краткосрочный период времени – *рывки* – моменты времени, когда цена инструмента существенно отклонялась от тренда в течение 10 минут после этого момента.

Что касается текстовой коллекции, то после стандартной предобработки каждый заголовок представлялся в виде мешка слов. Для каждого заголовка рассматривались как униграммы, так и биграммы. После фильтрации по частоте встречаемости слов осталось порядка 2000000 новостных заголовков, содержащих порядка 30000 уникальных униграмм и 25000 уникальных биграмм.

**Агрегация новостных заголовков** Для построения традиционных тематических моделей была произведения агрегация новостных заголовков по 20-минутным окнам; соответствующие счетчики униграмм и биграмм суммировались. Однако при этом нужно было аккумулировать метки классов. Для этого использовался следующий алгоритм:

- Если инструмент торговался в течение соответствующего отрезка времени, и за это время обнаружен хотя бы один рывок, то присваивалась метка 1;
- Если инструмент торговался в течение соответствующего отрезка времени, но рывков замечено не было, то присваивалась метка 0;
- Если инструмент не торговался в течение соответствующего отрезка времени, то пара (документ, инструмент) оставалась без метки.

## 4.2 Разделение выборки на обучение и контроль

Поскольку каждый объект выборки имеет временную метку, то имеет смысл разбивать выборку в зависимости от временных меток. В данном исследовании первые 70% объектов по времени попадают в обучение, следующие 20% используются для настройки гиперпараметров, а на последних 10% мы оцениваем итоговое качество.

| | LR | BTM + LR | WNTM + LR |
|---|---|---|---|
| AUROC | **0.60327** | 0.54872 | 0.56346 |
| AUC-PR | **0.75103** | 0.70921 | 0.71416 |

Таблица 1: Значение AUROC и AUC-PR на контрольной выборке для классификации рывка по любому инструменту

### 4.3 Прогнозирование рывков с использованием новостных заголовков

В рамках этого эксперимента производится оценивание качества прогнозирования с использованием тематических моделей, построенных на исходных новостных заголовках.

В качестве меток классов используется агрегация всех меток по инструментам: новости присваивается положительная метка, если в этот момент времени обнаружен рывок хотя бы по одному инструменту, в противном случае присваивается отрицательная метка.

Для оценивания качества прогнозирования используются следующие модели классификации:

- $l_2$-регуляризованная логистическая регрессия, в которой признаками являются счетчики униграмм и биграмм (LR); описание можно найти в [1];
- $l_2$-регуляризованная логистическая регрессия, в которой признаками являются тематические профили документов, построенные с использованием модели Biterm Topic Model (BTM + LR);
- $l_2$-регуляризованная логистическая регрессия, в которой признаками являются тематические профили документов, построенные с использованием модели Word Network Topic Model (WNTM + LR);

Коэффициенты $l_2$-регуляризации логистической регрессии для всех моделей настраиваются на отложенной выборке.

В табл. 1 показаны значения интегральных критериев качества AUROC и AUC-PR на тестовой выборке после обучения моделей с настроенными гиперпараметрами.

**Выводы** На основании результатов эксперимента мы видим, что при использовании тематических моделей, построенных на новостных заголовках, качество предсказания хуже по сравнению с популярными методами бинарной классификации, обученными на тех же данных. Из этого можно сделать несколько выводов:

- При поиске рывков по нескольким инструментам одновременно тематическая модель, построенная на новостных заголовках, может не дать выигрыша по сравне-

нию с другимим методами; возможно, имеет смысл рассматривать индивидуальные рывки по каждому инструменту;

Также такие в целом неутешительные результаты могут быть следствием того, что мы используем новостные заголовки, по которым даже эксперт не всегда может однозначно предсказать движение.

## 4.4 Прогнозирование рывков на агрегированных данных

В рамках данной серии экспериментов для оценивания качества прогнозирования с использованием тематических моделей использовались агрегированные новостные заголовки, а также преобразованные соответствующим образом метки классов. При оценке качества прогнозирования тематические модели классификации сравниваются с моделью логистической регрессии, где в качестве признаков используются счетчики слов в документах.

**Предсказание меток для одного инструмента** Эксперимент позволяет сравнить качество моделей при прогнозировании рывков для одного инструмента. Для этого используем инструмент EURUSD.

Параметры эксперимента:

- Коэффициент $l_2$-регуляризации логистической регрессии настраивается на отложенной выборке;
- Число итераций EM-алгоритма - 20;
- Коэффициенты модальностей для мультимодальной TM: униграммы – $\tau_{uni} = 1$, биграммы – $\tau_{bi} = 0.1$, метки $\tau_{EURUSD} = 1$.
- Коэффициент регуляризатора декоррелирования в зависимости от номера итерации $i$: $\tau_{1i} = 10^5$;
- Коэффициент регуляризатора сглаживания матрицы $\Phi$ ($\beta_0 = 1$, $\beta_w = 1/|W|$) в зависимости от номера итерации $i$: $\tau_{2i} = 0.4[i > 10]$ (сглаживание начинаем применять после 10 итераций);
- Коэффициент регуляризатора разреживания матрицы $\Theta$ ($\alpha_0 = 1$, $\alpha_w = 1/|T|$) в зависимости от номера итерации $i$: $\tau_{3i} = 0.2[i > 15]$ (сглаживание начинаем применять после 15 итераций);

В табл. 2 показаны значения метрик качества AUROC и AUC-PR на контрольной выборке. По результатам данного эксперимента можно сделать вывод, что использование тематической модели классификации существенно повышает качество классификации даже при небольшом числе тем; при увеличении числа тем значения метрик качества стабилизируются и остаются на примерно одинаковом уровне.

| | LR | TM | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\|T\| = 50$ | $\|T\| = 100$ | $\|T\| = 200$ | $\|T\| = 300$ | $\|T\| = 400$ | $\|T\| = 500$ |
| AUROC | 0.72941 | 0.76875 | **0.78126** | 0.77350 | 0.77703 | 0.77497 | 0.77480 |
| AUC-PR | 0.64263 | 0.66993 | **0.67957** | 0.66780 | 0.67240 | 0.66542 | 0.66550 |

Таблица 2: Значение AUROC и AUC-PR на контрольной выборке для классификации одного инструмента. Обозначение: LR-логистическая регрессия, TM-мультимодальная тематическая модель.

| | LR | TM | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\|T\| = 50$ | $\|T\| = 100$ | $\|T\| = 200$ | $\|T\| = 300$ | $\|T\| = 400$ | $\|T\| = 500$ |
| AUDUSD | 0.69413 | 0.69552 | **0.70142** | 0.69084 | 0.67869 | 0.69633 | 0.69667 |
| Dax | **0.73257** | 0.71452 | 0.71527 | 0.70783 | 0.70530 | 0.69880 | 0.71168 |
| EURUSD | 0.72941 | 0.76908 | **0.78163** | 0.77370 | 0.77697 | 0.77512 | 0.77499 |
| Eurostoxx | 0.70590 | 0.72836 | 0.72741 | 0.74126 | 0.74443 | 0.72858 | **0.74935** |
| GBPUSD | 0.70381 | 0.75676 | 0.77421 | 0.76019 | 0.76430 | 0.76435 | **0.78142** |
| MXNUSD | 0.76711 | 0.78453 | **0.80171** | 0.79659 | 0.79873 | 0.78629 | 0.77958 |
| NZDUSD | 0.65277 | 0.67754 | **0.68583** | 0.68090 | 0.66779 | 0.67872 | 0.68395 |
| OilBrentIce | **0.79917** | 0.78630 | 0.79468 | 0.78251 | 0.77263 | 0.77485 | 0.78314 |

Таблица 3: Значение AUROC на контрольной выборке для классификации всех инструментов. Обозначение: LR-логистическая регрессия, TM-мультимодальная тематическая модель.

**Задача предсказания меток для набора инструментов одновременно** В рамках данного эксперимента исследовалась способность тематической модели учитывать информацию о нескольких инструментах одновременно. Для этого мультимодальная тематическая модель, построенная с учетом всех инструментов одновременно, сравнивается с набором моделей логистической регрессии, построенных по каждому инструменту отдельно.

Параметры эксперимента совпадают с параметрами предыдущего, с учетом того, что в текущей TM классификации присутствуют модальности для каждого инструмента, а не только для EURUSD.

В табл. 3 и табл. 4 показаны результаты эксперимента. На основании этих результатов, а также результатов предыдущего эксперимента можно сделать следующие выводы:

- Тематическая модель с базовыми параметрами, построенная на всех сигналах сразу, имеет сопоставимое качество для всех инструментов по сравнению с набором моделей логистических регрессий, построенных для каждого инструмента по отдельности;

| | LR | TM | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\|T\| = 50$ | $\|T\| = 100$ | $\|T\| = 200$ | $\|T\| = 300$ | $\|T\| = 400$ | $\|T\| = 500$ |
| AUDUSD | **0.50369** | 0.44258 | 0.42589 | 0.40968 | 0.39747 | 0.41039 | 0.40226 |
| Dax | **0.49790** | 0.47368 | 0.47166 | 0.46042 | 0.47250 | 0.44193 | 0.45898 |
| EURUSD | 0.64263 | 0.66962 | **0.67917** | 0.66758 | 0.67229 | 0.66507 | 0.66562 |
| Eurostoxx | **0.36319** | 0.35076 | 0.33739 | 0.34834 | 0.35582 | 0.31911 | 0.35937 |
| GBPUSD | **0.51237** | 0.48604 | 0.50285 | 0.47702 | 0.48863 | 0.46901 | 0.50599 |
| MXNUSD | **0.43589** | 0.38995 | 0.40628 | 0.41247 | 0.43114 | 0.38818 | 0.38875 |
| NZDUSD | 0.54197 | **0.54460** | 0.52864 | 0.52376 | 0.50839 | 0.50938 | 0.52734 |
| OilBrentIce | **0.67183** | 0.61115 | 0.61731 | 0.60608 | 0.59345 | 0.56960 | 0.58646 |

Таблица 4: Значение AUC-PR на контрольной выборке для классификации всех инструментов. Обозначение: LR-логистическая регрессия, TM-мультимодальная тематическая модель.

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| peer snapshot | donald trump | bhp billiton | resumed trading | boe carney |
| inc peer | trump win | billiton say | due volatility | bank england |
| corporation peer | trump victory | centerra gold | paused due | central bank |
| company peer | hillary clinton | bhp say | volatility trading | carney say |

Таблица 5: Примеры интерпретируемых тем

- При добавлении в тематическую модель, настроенную на конкретный инструмент, меток других инструментов качество предсказания меток исходного инструмента не ухудшается. Это может быть косвенным свидетельством того факта, что на движение котировок финансовых инструментов могу влиять одни и те же события.

## 4.5 Интерпретируемость тем

Данный эксперимент проводится для проверки гипотезы о том, что тематические модели классификации, построенные для прогнозирования рывков, содержат хорошо интерпретируемые темы, влияющие на движение цен по мнению модели. Для этого используется тематическая модель из предыдущего эксперимента. При поиске интерпретируемых тем будем смотреть на биграммы.

В табл. 5 приведены примеры интерпретируемых тем среди тем, которые считаются моделью как влияющие на хотя бы один инструмент (т.е. $p(1|t) > 0$). Некоторые темы заслуживают отдельного комментария.

- Слово *peer* имеет в английском языке несколько значений, в том числе есть компания с названием PEER (Pure Energy & Environmental Resources). Поскольку

среди списка биграмм встречается *inc peer*, то логично предположить, что данная тема связана именно с этой компанией, которая в заголовках финансовых новостей может фигурировать как *Peer, Inc.*;

- Биграммы *bhp billiton* и *centerra gold* относятся к компаниям, связанным с добычей полезных ископаемых: BHP Billiton и Centerra Gold соответственно. Скорее всего, тема, объединяющая это биграммы, отвечает за новости, связанные с полезными ископаемыми;

- Сокращенное название банка Англии - BoE (Bank of England), а Mark Garney занимает в нем самую главную должность (en. Governor of the Bank of England). С июля по декабрь 2016 года наибольший интерес к банку Англии был во время споров о выходе страны из Евросоюза (en. Brexit) в конце июня – начале июля;

- Тема, связанная с победой Дональда Трампа на выборах Президента США в конце 2016 года, является хорошо интерпретируемой, и наличие выделенных биграмм в новостях действительно сильно коррелировало с движениями рынка. Но при этом данная тема является специфичной для конкретного датасета, и при рассмотрении новостей за другой период времени мы вряд ли будем наблюдать аналогичную тему.

Анализируя результаты, можно сделать вывод, что тематическая модель классификации позволяет выделять хорошо интерпретируемые темы, причем наличие слов из данных тем является существенным признаком наличия рывка.

# 5 Результаты, выносимые на защиту

- Тематическое моделирование позволяет строить модели на основе агрегатов новостных заголовков для объяснения движения различных финансовых инструментов за продолжительный промежуток времени.

- Тематические модели классификации имеют качество не хуже, чем модель логистической регрессии, однако при этом позволяют прогнозировать движение набора инструментов одновременно.

- Получаемые темы являются хорошо интерпретируемыми и помогают анализировать влияние событий на финансовые рынки.

# Список литературы


[1] *Bishop, C.* Pattern Recognition and Machine Learning / C. Bishop. — Springer, 2006.

[2] *Blei, D. M.* Latent dirichlet allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // *J. Mach. Learn. Res.* — mar 2003. — Vol. 3. — Pp. 993–1022.

[3] *Bollen, J.* Twitter mood predicts the stock market. / J. Bollen, H. Mao, X.-J. Zeng // *J. Comput. Science.* — 2011. — Vol. 2, no. 1. — Pp. 1–8.

[4] *Doyle, G.* Financial topic models / G. Doyle, C. Elkan. — 2009.

[5] *Fama, E. F.* Random walks in stock market prices / E. F. Fama // *Financial Analysts Journal.* — 1965. — Vol. 21, no. 5. — Pp. 55–59.

[6] *Fama, E. F.* Efficient capital markets: A review of theory and empirical work / E. F. Fama // *The Journal of Finance.* — 1970. — Vol. 25, no. 2. — Pp. 383–417.

[7] *Hisano, R.* High quality topic extraction from business news explains abnormal financial market volatility / R. Hisano, D. Sornette, T. Mizuno, T. Ohnishi, T. Watanabe // *CoRR.* — 2012. — Vol. abs/1210.6321.

[8] *Hofmann, T.* Probabilistic latent semantic analysis // Proc. of Uncertainty in Artificial Intelligence, UAI'99. — Stockholm: 1999.

[9] *Lo, A.* Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis / A. Lo // *Journal of Investment Consulting.* — 2005. — Vol. 7.

[10] *Ming, F.* Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization // 2014 IEEE International Conference on Data Mining. — IEEE, 2014. — Pp. 430–439.

[11] *Nassirtoussi, A. K.* Text mining for market prediction: A systematic review. / A. K. Nassirtoussi, S. R. Aghabozorgi, Y. W. Teh, D. C. L. Ngo // *Expert Syst. Appl.* — 2014. — Vol. 41, no. 16. — Pp. 7653–7670.

[12] *Nguyen, T. H.* Sentiment analysis on social media for stock movement prediction. / T. H. Nguyen, K. Shirai, J. Velcin // *Expert Syst. Appl.* — 2015. — Vol. 42, no. 24. — Pp. 9603–9611.

[13] *Vorontsov, K.* BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. // AIST / Ed. by M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets. — Vol. 542 of *Communications in Computer and Information Science.* — Springer, 2015. — Pp. 370–381.

[14] *Vorontsov, K.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization / K. Vorontsov, A. Potapenko // Analysis of Images, Social Networks and Texts / edited byD. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, R. E. Yavorsky. — Springer International Publishing, 2014. — Vol. 436 of *Communications in Computer and Information Science*. — Pp. 29–46.

[15] *Yan, X.* A biterm topic model for short texts. // WWW / Ed. by D. Schwabe, V. A. F. Almeida, H. Glaser, R. A. Baeza-Yates, S. B. Moon. — International World Wide Web Conferences Steering Committee / ACM, 2013. — P. 1445–1456.

[16] *Zuo, Y.* Word network topic model: a simple but general solution for short and imbalanced texts. / Y. Zuo, J. Zhao, K. Xu // *Knowl. Inf. Syst.* — 2016. — Vol. 48, no. 2. — P. 379–398.