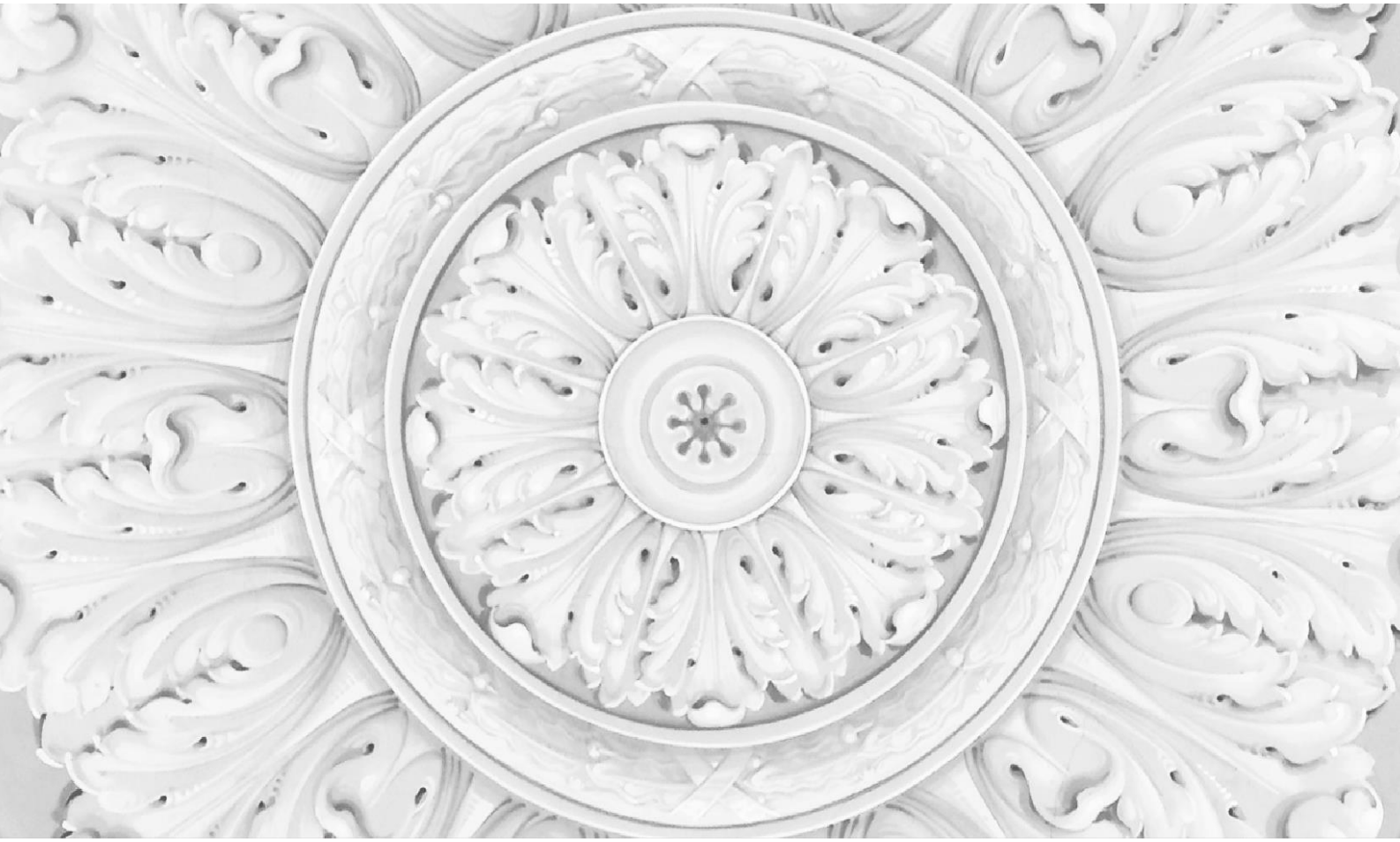
Банк России

Центральный банк Российской Федерации



СЕРИЯ ДОКЛАДОВ ОБ ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЯХ

Ксения Яковлева

Оценка экономической активности на основе текстового анализа

№25 / Октябрь 2017 г.

Ксения Яковлева
Банк России, Департамент исследований и прогнозирования
E-mail: YakovlevaKV@cbr.ru

Автор выражает благодарность Анне Кругловой, а также всем участникам семинаров в Банке России за ценные предложения и комментарии.

Адрес 107016, Москва, ул. Неглинная, 12
Телефоны +7 495 771-91-00, +7 495 621-64-65 (факс)
Сайт www.cbr.ru

Резюме

Данная работа посвящена описанию методики расчета высокочастотного индикатора, отражающего динамику экономической активности в стране. В качестве исходных данных используются новостные статьи, взятые из интернет-ресурсов. Анализ новостных статей производится методами текстового анализа и машинного обучения, которые разработаны сравнительно недавно, однако довольно быстро получили широкое распространение в научных работах, в том числе экономических. Это связано с тем, что новости являются не только основным источником информации, с их помощью можно также узнавать настроения журналистов и опрошенных респондентов относительно текущей ситуации и преобразовывать их в количественные данные.

Ключевые слова: оценивание экономической активности, текстовый анализ, машинное обучение.

JEL классификация: C51, C81, E37.

ВВЕДЕНИЕ

В последнее время особую популярность приобрело направление «большие данные» (Big Data) в связи с колоссальным ростом объемов цифровой информации. Этот рост обусловлен повсеместным распространением технологий и доступа к сети Интернет, что позволяет пользователям на постоянной основе создавать новую информацию. К такой информации относятся практически все данные сети Интернет, находящиеся в свободном доступе, а именно: сайты интернет-магазинов, сайты по поиску работы, новостные источники, социальные сети, блоги и многое другое. При этом значительная часть информации в сети Интернет представлена в неструктурированном виде, то есть в форме текста. Это не позволяет человеку самостоятельно обрабатывать большой объем информации, поэтому исследователями в разных сферах, в том числе экономистами, разрабатываются новые статистические подходы к извлечению и анализу неструктурированных интернет-данных.

Основным преимуществом данных сети по сравнению с обычными статистическими данными является их многообразие, возможность рассчитать показатель, не учитывающийся в официальной статистике. Важно также отметить, что интернет-данные обладают гораздо большей высокочастотностью и оперативностью по сравнению с официальными статистическими данными.

Целью данной работы является построение высокочастотного индикатора, рассчитанного на основе ежедневных новостей, для оценки динамики экономической активности в стране. Необходимость построения данного индикатора обусловлена отсутствием аналогичных показателей и запаздыванием официальных данных по экономической динамике. Например, основной показатель экономического роста – ВВП публикуется на квартальной основе с задержкой в 1–1,5 месяца после окончания квартала, что не позволяет оперативно отслеживать экономическую динамику и принимать соответствующие решения.

Вышесказанное демонстрирует высокую актуальность использования больших данных в экономике. Причем актуальность будет только расти в связи с большим спросом на данные исследования.

Тема «Большие данные в экономике» – новое направление, особенно для России: на сегодняшний день в нашей стране практически не существует работ по

данной тематике, что указывает на новизну этой темы. За основу был взят подход, описанный в статье A. Thorsrud (2016), но с некоторыми модификациями.

В разделе 1 настоящего исследования представлены методология модели и исходные данные, в разделе 2 приводятся результаты построенной модели.

1. ИСХОДНЫЕ ДАННЫЕ И СПЕЦИФИКАЦИЯ МОДЕЛИ

1.1. Данные

В работе используются два типа данных: неструктурированные и структурированные. В качестве неструктурированных данных, то есть данных, которые не имеют определенной структуры, выступают ежедневные новостные статьи, взятые из интернет-ресурса. Второй тип данных – это ежемесячный статистический показатель – композитный индекс деловой активности PMI (Purchasing Managers Index). Индекс деловой активности PMI используется в качестве прокси ВВП (в связи с недостаточно длинными временными рядами новостных статей[1]).

Новостные статьи были собраны с информационного ресурса, посвященного экономической тематике. Его выбор обусловлен широким охватом экономических новостей, отсутствием нерелевантных тем и простотой веб-скрапинга[2]. Общее количество статей составило около 50 000, совокупный объем слов – 20–25 млн, что является приемлемым для проведения анализа.

Данные по композитному индексу деловой активности PMI были взяты с сайта агентства Bloomberg.

Для построения индикатора использовалась временная выборка с января 2014 года по август 2017 года.

1.2. Модель

В построении новостного индекса можно выделить три основных этапа. Первый этап состоит из извлечения списка тем, содержащихся в новостных текстах. Второй этап определяет тональность новостных текстов, что позволяет разделить темы на положительные или отрицательные и отследить их динамику. Третий этап заключается в построении линейной регрессии, где в качестве зависимой переменной выступает индекс деловой активности PMI, а в качестве регрессоров – преобразованные с помощью метода главных компонент темы новостей.

[1] В связи с тем, что ВВП рассчитывается на ежеквартальной основе, для построения регрессионного анализа требуется более длинный временной новостной ряд.
[2] Технология, позволяющая автоматически извлекать и сохранять данные из интернет-ресурсов.

Прежде чем перейти к выполнению всех трех этапов построения индекса, необходимо подготовить данные, то есть преобразовать неструктурированный текст в структурированный вид. Подготовка данных является важным этапом в текстовом анализе, так как, во-первых, позволит уменьшить размерность данных, что значительно ускорит процесс обработки информации; во-вторых, лучшая подготовка текста в самом начале позволит получить более качественные и интерпретируемые темы в итоге.

Подготовка включает несколько шагов. Первым шагом является приведение всех слов к первоначальному виду – стемматизация. Для стемматизации использовалась программа MyStem – свободно распространяемый морфологический анализатор русского языка, созданный в Яндексе в 1997 году. Принципы работы описаны в статье одного из основателей данной поисковой системы, программиста И. Сегаловича. На втором шаге обработки текста происходит удаление пунктуации, чисел, лишних пробелов, «стоп-слов»[3]. Проведенная «фильтрация» новостных текстов позволяет значительно уменьшить исходные данные, не теряя при этом смысловую составляющую.

Преобразованные слова в «отфильтрованных» текстах называют *терминами* (англ. term – термин), на основе которых строится матрица dtm (document-term matrix), где каждая строка матрицы определяет отдельный термин, а каждый столбец – отдельный документ.

Используя матрицу dtm, можно перейти к первому этапу работы – выявлению тем в корпусе. Темы в работе были выявлены с помощью вероятностного тематического моделирования, которое определяет тему как множество слов, каждое из которых упорядочено по степени его принадлежности к теме.

Одним из наиболее популярных вероятностно тематических методов моделирования является метод латентного распределения Дирихле (latent Dirichlet allocation – LDA). Использование данного метода в текстовом анализе было предложено D. Blei, A. Ng, M. Jordan (2003), которые на основе модели, применяемой в информационно-поисковой системе, показали, что каждый документ имеет несколько тем, вероятностное распределение которых можно выявить методом LDA.

[3] «Стоп-слова» – это слова-связки, которые не несут смысловой нагрузки. К словам-связкам относятся союзы и союзные слова, местоимения, предлоги, частицы, междометия, указательные и вводные слова, а также ряд некоторых существительных, глаголов и наречий.

Латентное распределение Дирихле – трехуровневая иерархическая байесовская модель, в которой каждый документ в корпусе смоделирован как совокупность ненаблюдаемых, скрытых тем. Каждое слово в текстовом документе, согласно LDA, принадлежит неизвестной нам теме, а каждая тема моделируется из изначально заданных вероятностей тем:

$$p(\theta, z, w|\, \alpha, \beta) = \; p(\theta \mid \alpha) \; \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta) \quad ,$$

где α и β – задаваемые векторы – параметры распределения Дирихле;

$\theta \sim \mathrm{Dir}(\alpha)$ – распределение тем в каждом документе;

$z_n \sim \mathrm{Dir}(\beta)$ – распределение слов в каждой теме;

w_n – соответствие слов в документе темам.

На выходе модели LDA получаются векторы, показывающие, как распределены темы в каждом документе, и распределения, показывающие, какие слова более вероятны для каждой темы. Результатом модели LDA является список слов, наиболее характерных для каждой темы. Список слов ранжируется по мере убывания, при этом из него извлекаются первые пять слов, на основе которых в последующем происходит расчет частоты встречаемости тем.

Для получения данных на ежедневной основе все статьи за день суммируются, а затем на основе пяти первых слов каждой темы вычисляется их частота за день. Полученные данные показывают частоту упоминания каждой темы за день, однако они не отражают тона темы (положительный или отрицательный). Поэтому вторым этапом в построении индекса является определение тона текста.

В литературе существует четыре основных подхода к определению тональности:

1) подходы, основанные на правилах;
2) подходы, основанные на словарях;
3) машинное обучение с учителем;
4) машинное обучение без учителя.

A. Thorsrud (2016), чья работа была взята за основу при построении данного индекса, определил тональность текста с помощью подхода, основанного на словаре. Для идентификации положительного и отрицательного тона автор использовал словарь Harvard IV-4 Psychological Dictionary, в котором уже

определены позитивные и негативные слова. Однако данный словарь не отражает специфичности экономической терминологии и русского языка. Поэтому для определения тональности текста в данной работе был использован подход, основанный на машинном обучении с учителем (supervised learning), так как он, как правило, демонстрирует высокое качество классификации. В качестве метода был выбран метод опорных векторов (support vector machine – SVM), который позволяет разделить выборку на два класса с помощью разделяющей гиперплоскости, для того чтобы расстояние от гиперплоскости до ближайших точек множества было максимальным.

Однако для применения данного метода необходима обучающая коллекция, то есть уже классифицированные тексты («позитивный текст» и «негативный текст») аналогичного формата. Поэтому на первом шаге была построена обучающая коллекция. Все новости в обучающей коллекции были вручную разделены на два класса – «позитивный» и «негативный». На втором шаге обученная модель была применена в отношении всего оцениваемого ряда.

Вместе с тем не все тексты можно однозначно классифицировать как позитивные или негативные, часть из них может быть написана в нейтральном тоне. Результатом моделирования SVM является класс тональности («+1» или «-1») и вероятность принадлежности текста к какому-либо классу тональности. Следовательно, если модель определяла тональность текста с вероятностью ниже 60%, то его тон принимался за нейтральный. Тексты, приравненные к нейтральному классу, были исключены из анализа, так как они не несут необходимой нам информации.

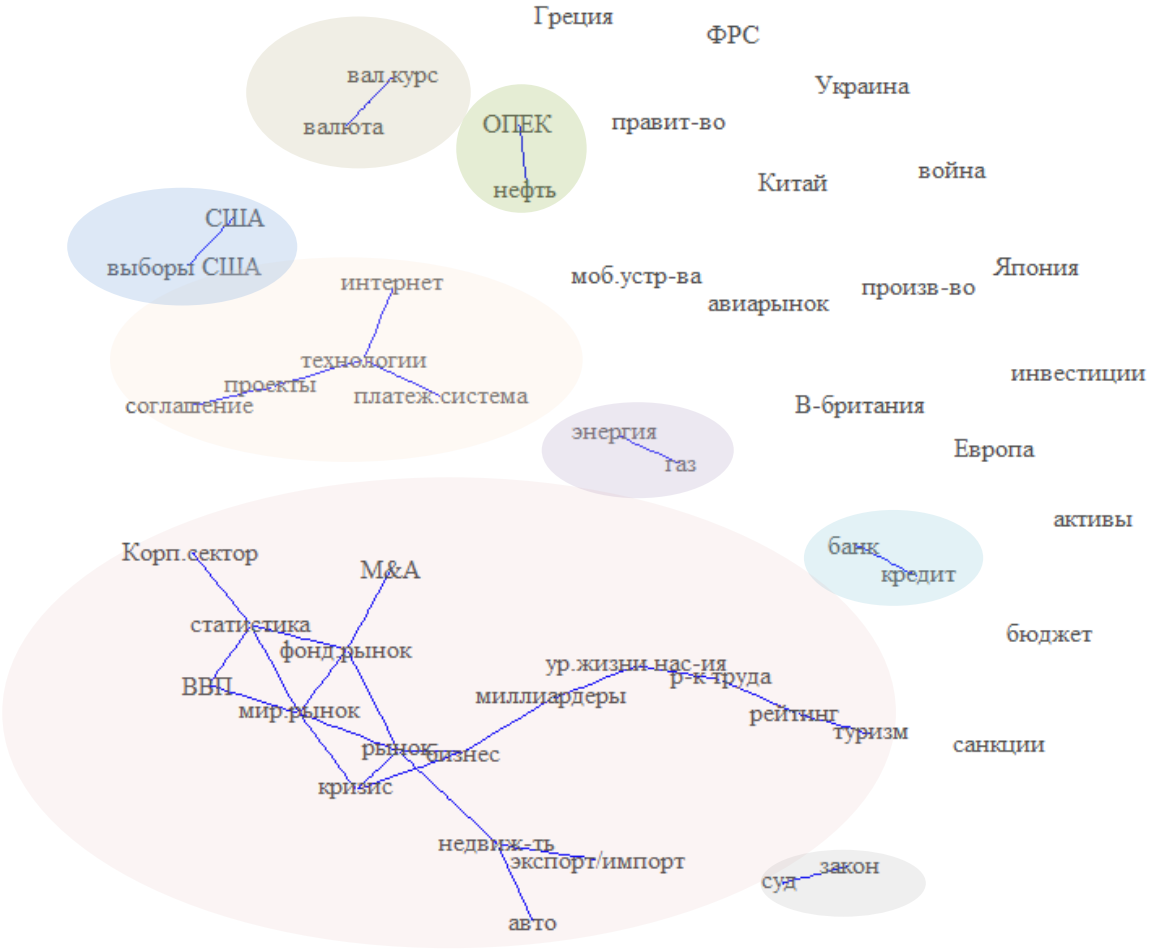
В результате все темы, полученные на первом этапе, были скорректированы на рассчитанную тональность и стали регрессорами в эконометрической модели, построенной на третьем этапе. В качестве эконометрической модели была использована классическая линейная модель множественной регрессии.

Для уменьшения размерности данных был использован метод главных компонент (Principal Component Analysis – PCA). Выбор PCA связан с тем, что он позволяет сократить количество регрессоров, потеряв наименьшее количество информации.

2. РЕЗУЛЬТАТЫ АНАЛИЗА

Согласно модели LDA, было выявлено 50 тем, обеспечивающих наилучшее статистическое разложение корпуса. Модель LDA не присваивает темам имена, однако увидев наиболее часто встречающиеся слова в каждой теме, мы можем понять, о чем данная тема, и присвоить ей адекватное название. Например, с января 2014 года по январь 2017 года основными темами в новостных статьях были темы, связанные с валютным курсом, нефтью, банковским сектором, ситуацией в США и так далее (Рисунок 1).

Рисунок 1. Темы, выявленные с помощью метода LDA, и связь между ними



Источник: расчеты автора.

На рисунке представлена также сетевая визуализация тем, где были удалены все ребра, которые имеют корреляцию менее 20%. Для расчета корреляции использовалась матрица, где в качестве строк были взяты 50 тем, а в качестве столбцов – все термины, употребляемые новостным сайтом за рассматриваемый период. В итоге получилась матрица размером 50x243577, где на пересечении строк и столбцов была рассчитана вероятность появления каждого слова (термина) в каждой теме. Затем на основе вероятностей была рассчитана корреляционная матрица между темами размерностью 50x50.

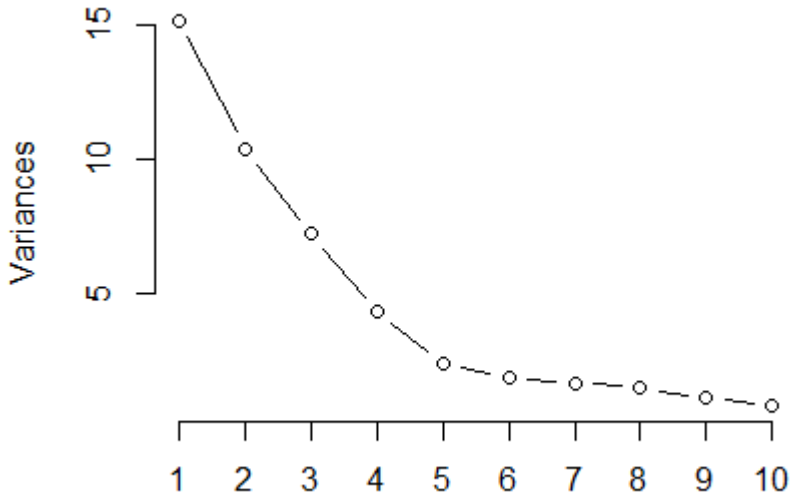
Согласно рисунку, половина тем разделилась на несколько кластеров – как крупных, так и небольших. Данная кластеризация не будет использована при дальнейшем анализе, однако является удобным способом для визуализации рассматриваемых тем.

В итоге на основе 50 выявленных тем было рассчитано 50 ежедневных временных рядов, которые в дальнейшем были скорректированы на тональность, полученную с помощью метода SVM. Для оценки качества работы метода SVM обучающая выборка была разбита на две части. На основе первой части (9/10 от обучающихся данных) была построена модель, а на основе второй части (1/10 от обучающихся данных) была проверена точность работы алгоритма. Точность была рассчитана как доля верно предсказанных значений от всех значений и составила 68%, что является достаточно неплохим результатом.

Для устранения дневного шума во временном ряду каждой темы данные были сглажены скользящей средней за 80 дней. Для сопоставления ежедневных тем с месячным индексом деловой активности PMI темы были преобразованы в месячные путем нахождения среднего значения за месяц. В результате получилось 50 месячных временных рядов-регрессоров, каждый из которых характеризует определенную тему.

Для построения линейной регрессии был использован метод PCA, который позволил уменьшить размерность регрессоров (Рисунок 2).

Рисунок 2. Темы, преобразованные с помощью метода PCA



Источник: расчеты автора.

Из 50 изначально заданных тем осталось четыре регрессора, хорошо описывающих зависимую переменную – показатель экономической активности PMI. Регрессионный анализ показал, что первая компонента не является значимой в построенном уравнении, в то время как компоненты со второй по пятую значимы и объясняют 85% регрессии (Таблица 1).

Таблица 1. Регрессия на индекс деловой активности PMI

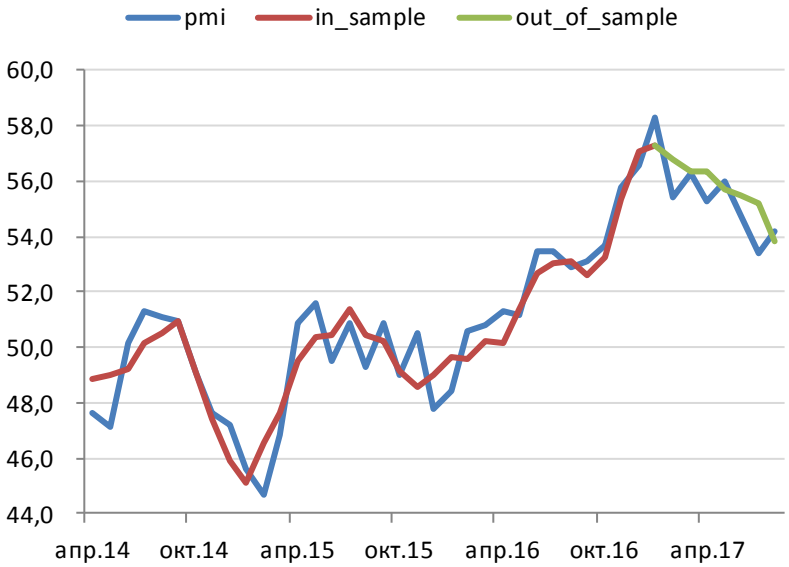
Зависимая переменная = индекс PMI		
Переменная	β	t-статистика
PCA(2)	0,2329	10,28 ***
PCA(3)	-0,1054	-3,87 ***
PCA(4)	-0,1268	-3,60 **
PCA(5)	-0,2368	-5,00 ***
F-статистика	39,67 ***	
R 2 – adjusted	0,8455	

*Примечание: ***; ** и * – значимость оценки коэффициента на 0,1; 1 и 5% соответственно.*

Для верификации качества построенной линейной регрессии выборка была разбита на обучающую (in sample) и контрольную (out of sample). Обучающая выборка взята за период с января 2014 года по январь 2017 года, контрольная – с февраля 2017 года по август 2017 года. Сравнение контрольной выборки с

фактическими данными PMI говорит о достаточно неплохой прогнозной силе модели (Рисунок 3).

Рисунок 3. PMI и рассчитанный индекс на основе новостей



Источники: IHS Markit, расчеты автора.

Для оценки качества построенной модели также была использована средняя абсолютная ошибка (MAE). Для данного прогноза значение MAE составило 0,81%. При этом средняя абсолютная ошибка прогнозирования при использовании другой модели – модели автокорреляции первого порядка AR(1) – составила 2,7% (Таблица 2).

Таблица 2. Средняя абсолютная ошибка прогнозной модели

	Средняя абсолютная ошибка (MAE)
Модель, построенная на основе новостей	0.81
Модель AR(1)	2.7

ЗАКЛЮЧЕНИЕ

В работе была представлена модель, оценивающая экономическую динамику на основе новостных статей. Приведенные расчеты показали, что использование такой неструктурированной информации, как новости, является не менее важной составляющей при прогнозировании экономической активности, чем использование обычных статистических показателей.

Разработанная методика достаточно успешно справилась с решением задачи прогнозирования экономической динамики, о чем свидетельствуют полученные оценки качества модели. Это позволяет сделать вывод о том, что новостные данные обладают достаточно хорошей прогнозной силой. С помощью разработанного новостного индекса можно отслеживать динамику не только экономической активности на ежедневной основе, но также разрабатывать иные индикаторы, что позволит более оперативно реагировать на текущую экономическую ситуацию и принимать соответствующие решения.

ЛИТЕРАТУРА

1. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. January, 2003.
2. Bloom N., Baker S., Davis S. Measuring Economic Policy Uncertainty // THE QUARTERLY JOURNAL OF ECONOMICS. November, 2016.
3. Doms M., Morin N. Consumer Sentiment, the Economy, and the News Media // Finance and Economics Discussion Series (FEDS). September, 2004.
4. Kholodilin K., Thomas T., Ulbricht D. Do media data help to predict German industrial production? // Journal of Forecasting. 2017.
5. Shapiro A., Sudhoh M., Wilson D. Measuring News Sentiment // FEDERAL RESERVE BANK OF SAN FRANCISCO WORKING PAPER SERIES. January, 2017.
6. Thorsrud A. Words are the new numbers: A newsy coincident index of business cycles // Norges Bank Research. Working Paper. February, 2016.
7. Голощапова И., Андреев М. Оценка инфляционных ожиданий российского населения методами машинного обучения // Вопросы экономики, июнь 2017 г., № 6.