Банк России
Центральный банк Российской Федерации

# Тексты экономических новостей: полезное дополнение к официальной статистике?

**Аналитическая записка**
**Департамента исследований**
**и прогнозирования**
**Банка России**

Ноябрь 2018

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

***РЕЗЮМЕ***

*Динамично развивающийся современный мир, постоянное стремительное увеличение поступающей информации предопределяют неуклонно растущий интерес к экономическим исследованиям, которые основываются на анализе больших массивов неструктурированной информации. Очевидно, что объем такой информации в разы превосходит массивы более привычных экономистам структурированных данных и является сложным для технической обработки. Но при этом он открывает принципиально новый спектр возможностей для исследователей.*

*На фоне повышенного внимания экономистов к анализу и обработке неструктурированной информации растет, в частности, и научный интерес к анализу всевозможных текстовых документов с помощью автоматической обработки текста.*

*Отталкиваясь от опубликованной ранее методики построения индекса экономической активности в России на основе текстов экономических новостей, мы попытаемся вкратце осветить основные особенности работы с текстовыми данными и дать ответ на вопрос о потенциальной сфере практического применения такого показателя.*

*Проведенный нами анализ продемонстрировал, что новостные статьи, заимствованные из интернет-источников, могут достаточно точно предсказать основные краткосрочные тренды в динамике деловой активности. Однако новостной индекс при этом может рассматриваться как полноценный и самодостаточный индикатор состояния экономики, который способен оперативно «улавливать» полезную информацию, не содержащуюся в показателях официальной статистики и опросных данных. С помощью предложенной методики разработанного новостного индекса также можно разрабатывать и иные аналитические индикаторы для анализа текущей экономической ситуации, которые могут быть полезными в том числе при проведении денежно-кредитной политики центральным банком.*

Текстовый анализ становится все более популярным в научном и профессиональном мире. Это объясняется тем, что всевозможная информация о любых фактах, наблюдениях, событиях содержится главным образом в неструктурированном формате[1]. Согласно некоторым исследованиям, компании хранят в неструктурированном виде около 80–90% от всей имеющейся у них информации, и объем этой информации стремительно увеличивается[2].

На фоне этого растет научный интерес к анализу всевозможных текстовых документов с помощью автоматической обработки текста, что ведет к появлению большого количества исследовательских работ в этой области в последнее время. Приведем лишь несколько примеров. В сфере маркетинга исследуются отзывы покупателей, представленные в текстовой форме. В финансовой области тексты из

[1] Неструктурированная информация – это информация, которая либо не имеет заранее определенной структуры данных, либо не организована в установленном порядке.

[2] Kambies T., Roma P. Dark analytics: illuminating opportunities hidden within unstructured data. Tech Trends 2017. URL: https://www2.deloitte.com/insights/us/en/focus/tech-trends/2017/dark-data-analyzing-unstructured-data.html (дата обращения – 02.10.2018).

финансовых новостей, социальных сетей используются для прогнозирования движения цен активов. В макроэкономике тексты используется для прогнозирования колебаний инфляции, экономического роста, безработицы и других показателей. При этом текстовый анализ (*text mining*) в экономической сфере представляется той областью, где еще остается обширное пространство для исследований.

Текстовый анализ позволяет решать множество задач, которые включают в себя классификацию, кластеризацию, извлечение мнений и фактов, построение экспертных и вопросно-ответных систем, поиск по ключевым словам и многое другое. Для экономистов дополнительная информация из неструктурированных массивов данных может быть полезна с точки зрения более подробного изучения свойств различных экономических показателей и явлений, а также их прогнозирования. Подобный анализ важен и для центральных банков, в том числе для их денежно-кредитной политики (Bholat D., Hansen S., 2015).

Центральные банки разных стран уже внедрили методы текстового анализа в исследовательскую работу. Основными источниками информации при проведении таких исследований выступают новостные статьи, социальные сети, различные отчеты, выступления официальных лиц и иная информация.

Если обращаться непосредственно к недавним примерам, то текстовый анализ был использован Банком Англии, чтобы выявить гетерогенность на рынке труда и его влияние на выпуск и производительность. В частности, в одном из недавних исследований Turrell A. и Speigner B. (2018) использовали онлайн-вакансии, взятые с сайта по поиску работы, и оценили, как повысится выпуск и производительность труда, если несоответствия между спросом и предложением на рынке труда по занятиям и регионам будут устранены. В Банке Канады провели исследование, чтобы выявить, как официальные сообщения регулятора влияют на динамику доходностей и волатильности краткосрочных и долгосрочных процентных ставок (Hendry S, Madeley A., 2010). Во многих экономических исследованиях, посвященных текстовому анализу, предлагаются модели для краткосрочной оценки *(nowcast*) разнообразных экономических показателей. Основной вклад таких работ заключается в улучшении качества краткосрочных оценок и прогнозов наблюдаемых макроэкономических показателей с помощью новостной базы (Ardia D., Bluteau K., 2017, Doms M., Morin N., 2004, Nyman R., Ormerod P., 2014, Shapiro A., Sudhoh M., 2017, Bloom N., Baker S., Davis S., 2016). Например, в Банке Норвегии был построен индекс бизнес-цикла на основе текстов ежедневных деловых газет, который каждый день оценивает динамику ВВП для заданного квартала (Thorsrud A., 2016).

В целом же научные публикации или как минимум ссылки на результаты применения текстового анализа в аналитических материалах, как исследовательских институтов, так и непосредственно центральных банков, можно находить едва ли не еженедельно.

Предложенная в ходе нашего исследования модель текстового анализа[3] позволяет получить оценку месячного композитного индекса деловой активности (PMI[4]) на ежедневной основе с помощью новостных экономических статей. Построение высокочастотного (ежедневного) индекса, в свою очередь, может позволить своевременно отслеживать в краткосрочных данных важные тенденции, которые требуют оперативной реакции мерами политики.

В качестве текстовой информации нами использовались ежедневно публикуемые экономические новости. Выбор новостей связан с тем, что они отражают все события, происходящие как внутри страны, так и за рубежом, влияют на настроения и поведение субъектов экономики, которые впоследствии принимают экономические решения.

При выборе новостного источника основными критериями стали:

1. соответствие новостей экономической тематике;

2. наличие в интернете большого объема новостной информации за до- статочно длительный период времени (как минимум 3-4 года);

3. простота веб-скрапинга[5], т.е. возможность легко и быстро извлечь ин- формацию с сайта.

С учетом описанных выше критериев нами был выбран новостной ресурс “Вести.Экономика (vestifinance.ru)”, посвященный исключительно обзору экономических событий, происходящих как в России, так и за рубежом. Следует сразу оговориться, что одним из направлений дальнейшего развития представленного исследования может стать возможное расширение числа используемых нами источников.

При построении новостного индекса Банка России использовалось более 60 тыс. статей за период 2014–2018 гг., представленных на экономическом новостном ресурсе. Количество статей варьируется от месяца к месяцу и в среднем составляет около 1100 (Рисунок 1). Статьи были собраны нами за период с 2014 г. по настоящее время.

[3] Более подробная информация в статье: Яковлева К. (2017). Оценка экономической активности на основе текстового анализа // Серия докладов об экономических исследованиях в Банке России, № 25.
[4] Purchasing Managers Index – макроэкономический показатель, характеризующий состояние экономики в производственном секторе и сфере услуг, рассчитанный по результатам опроса менеджеров.
[5] Веб-скрапинг (англ. web scraping) – технология, получения данных с интернет-страниц.

**Рисунок 1. Количество новостных статей в месяц, ед.**


*Источники: интернет-новости, расчеты ДИП.*

Для дальнейшего анализа статьи распределяются по темам. Каждая новостная статья в разной степени отражает одну или несколько тем. Чтобы это выявить, использовалось тематическое моделирование, позволяющее автоматически рассортировывать по темам все собранные статьи.

Базовыми тематическими моделями являются вероятностный латентный семантический анализ (*Probabilistic Latent Semantic Analysis, PLSA*) и модель латентного размещения Дирихле (*Latent Dirichlet Allocation, LDA*). Каждая из них имеет свои достоинства и недостатки, которые зависят прежде всего от особенностей обучения на исторической ретроспективе и лингвистической обработки (Воронцов К., 2013). Мы не будем подробно останавливаться на этих методологических аспектах. Отметим лишь, что на текущем этапе исследования по построению и анализу новостного индекса нами была использована модель LDA как вероятностная модель, наиболее распространенная среди аналогичных исследований для других стран.

Результатом применения модели LDA является список тем, выявленных в новостных статьях и представленных в виде наиболее характерных отдельных слов (*униграмм)* для каждой рассматриваемой темы (Blei D, 2003).

Проведенный нами анализ показал, что для всех собранных статей оптимальным количеством является 50 тем. При большем их числе темы начинают дублироваться, а при меньшем – несколько тем объединяются в одну. Для исследования

этих тем нами была применена модель LDA на предварительно обработанный *корпус*[6].

На рисунке изображено 100 униграмм одной из тем (Рисунок 2). При этом необязательно просматривать все эти слова, чтобы понять, к чему они относятся. Согласно исследованию Kholodilin K. и Thomas T. (2017), первые 10 слов содержат в себе 30% информации о теме, что является достаточным для ее полного понимания. Поэтому по наиболее выделяющимся словам, можно сразу понять, что *униграммы* относятся к бюджетной системе.

**Рисунок 2. Униграммы в одной из тем, выявленных с помощью тематического моделирования LDA**


*Источники: интернет-новости, расчеты ДИП.*

Список основных тем, которые были выявлены нами с помощью модели LDA, приведен в Таблице 1. *Тема 1* включает в себя такие слова, как «облигация», «кредит», «доходность», – они раскрывают область экономики, которая характеризует долговой рынок. Наиболее часто используемыми словами в *теме 2* являются «доллар», «рубль», «курс» – тема содержательно относится к валютному курсу. Таким образом, модель выдает нам только ключевые слова (*униграммы*), на основе которых мы самостоятельно должны «назвать» каждую из перечисленных тем.

**Таблица 1**

**Результаты тематического моделирования новостных статей**

| Тема | Ключевые слова (униграммы) |
|---|---|
| Тема 1 | Облигация, кредит, доходность, бумага, объем |
| Тема 2 | Доллар, рубль, курс, евро, валюта |
| Тема 3 | Украина, украинский, Крым, Киев, гривна |

[6] *Корпус* – это совокупность текстовых документов. Под «*предварительно обработанным корпусом*» подразумевается удаление всех бессмысленных с точки зрения машинного обучения слов, числовых значений, знаков препинания и иных символов, которые следует исключить из текста для фильтрации шума в документах.

| | |
|---|---|
| Тема 4 | Фонд, инвестор, актив, финансовый, инвестиция |
| Тема 5 | Экспорт, товар, продукция, импорт, тонна |
| Тема 6 | Банк, банковский, Сбербанк, капитал, ВТБ |
| Тема 7 | Долг, МВФ, кредитор, помощь, дефолт |
| Тема 8 | Газ, Газпром, поставка, куб, поток |
| Тема 9 | Месторождение, нефть, проект, Роснефть, добыча |
| Тема 10 | Развитие, проект, инвестиция, бизнес, создание |
| Тема 11 | Китай, китайский, КНР, юань, Азия |
| Тема 12 | Источник, энергия, энергетика, уголь, электроэнергия |
| Тема 13 | США, Трамп, американский, Обама, штат |
| Тема 14 | Нефть, цена, баррель, добыча, ОПЕК |
| Тема 15 | Минфин, бюджет, доход, расход, дефицит |
| Тема 16 | Ставка, ФРС, политика, банк, федрезерв |

*Источники: интернет-новости, расчеты ДИП.*

Среди слов всех тем можно заметить достаточно общие по значению слова, у которых появляется смысл только в определенном контексте. Это можно проследить по *теме 10,* в которой есть *униграмма* «создание». Смысл слова «создание» достаточно широкий, и оно может встречаться в нескольких темах, однако вероятность появления данного слова в других темах значительно ниже, поэтому слово «создание» в большей мере будет характеризовать именно тему 10, а не какую-нибудь другую.

Важно отметить, что темы, представленные в новостных статьях, меняются со временем. Их эволюция может быть рассмотрена с точки зрения двух аспектов. *Во-первых,* происходит замена одних тем другими. *Во-вторых,* изменяется их интенсивность, которая отражает степень интереса к ним со стороны СМИ.

Что касается замены одних тем другими, то в экономической сфере темы достаточно общие и быстро не эволюционируют (например, валютный курс, рынок труда, инфляция и другие). Мы вкратце коснемся этого аспекта в заключительной части записки, где обсуждаются результаты прогнозирования деловой активности в российской экономике с помощью новостного индекса. В рассматриваемых нами темах будет меняться только их интенсивность, то есть степень интереса к ним со стороны СМИ.

Чтобы проиллюстрировать, как может в целом эволюционировать интенсивность тем, рассмотрим темы «Великобритания», «США», «Валютный курс», использовав простой подсчет встречаемости слов в новостных статьях.

Интенсивность тем с января 2014 г. по июль 2018 г. в зависимости от рассматриваемого периода ощутимо снижается или, наоборот, возрастает. Всплеск количества новостей можно заметить в отношении темы «США» (Рисунок 3). Резкий скачок (более чем в два раза) вызван президентскими выборами в США, которые прошли в ноябре 2016 года. По теме «Великобритания» (Рисунок 5), можно проследить, в какие периоды времени Великобритания привлекала наибольшее внимание СМИ. Максимальное количество новостей приходится на июнь 2016 г., что связано с референдумом по поводу членства Великобритании в Европейском союзе, который состоялся 23 июня 2016 года.

**Рисунок 3. Тема «США»**

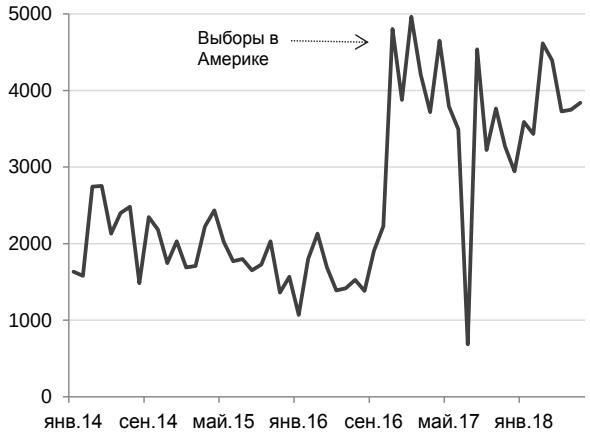
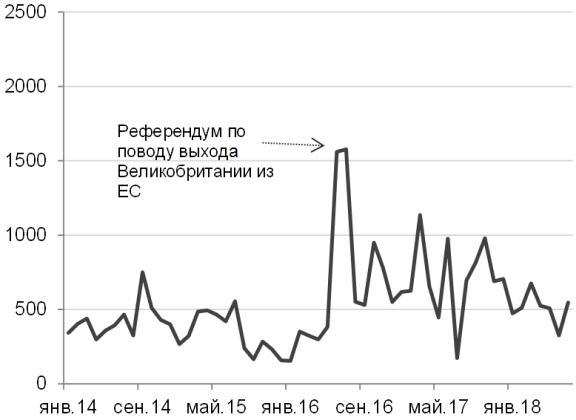



**Рисунок 4. Тема «Валютный курс»**


**Рисунок 5. Тема «Великобритания»**




*Источники: интернет-новости, расчеты ДИП.*

При этом новости оказывают разное влияние на российскую экономику. Например, референдум в Великобритании и расследование инцидента в Солсбери – примеры новостей по одной теме, которые могут оказаться разными по степени влияния на Россию. Для того чтобы учесть данный аспект, текстовый анализ предполагает корректировку новостей по тональности. В частности, в ходе наших расчетов новостного индекса мы присваивали новости значение *«1»*, если она положительно сказывается на России, и *«-1»* – если отрицательно.

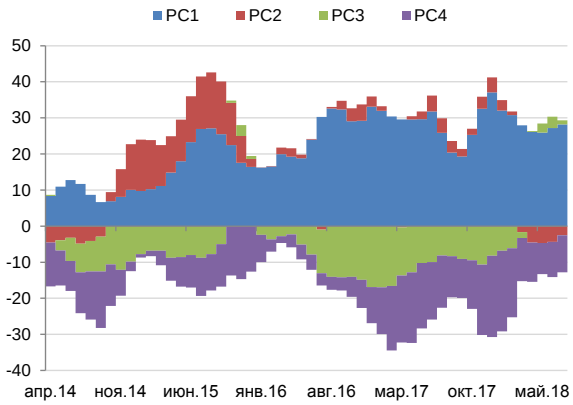
К теме «Валютный курс» (Рисунок 4) в свою очередь относятся такие *униграммы*, как «доллар», «рубль», «курс», «евро», «валюта», поэтому тема не будет хорошо отражать динамику официального курса той или иной валюты, а даст общую картину напряженности на валютном рынке. Например, можно заметить, что в конце 2014 г. динамика количества новостей коррелировала с динамикой официального курса рубля Банка России. Это объясняется тем, что обесценение рубля на фоне высокой геополитической неопределенности и снижения цен на нефть повлекло за собой большое количество релевантных публикаций в СМИ. Последующие эпизоды ослабления рубля на фоне происходящей постепенной адаптации экономики к ухудшению внешнеэкономических условий, которые сопровождались уже более умеренной реакцией субъектов экономики на курсовые колебания, не так интенсивно освещались журналистами.

При этом важно отметить, что классификация текстов с помощью *униграмм* может быть и ошибочной, поскольку одни и те же слова могут иметь разный смысл. Поэтому построение тематической модели с использованием других методов представляет собой актуальную исследовательскую задачу. Одним из таких методов, например, является использование *биграмм* – словосочетаний из двух слов, таких как «*ценная бумага*», *«валютный курс»* и прочие словосочетания. Однако использование *биграмм* может в определенных ситуациях усложнить работу и, наоборот, привести к худшему результату. В таком случае надо попытаться найти компромиссную модель, позволяющую, с одной стороны, хорошо интерпретировать смысловое значение текста и, с другой стороны, не усложнять модель.

Процедура, на базе которой строится наш новостной индекс, предполагает, что полученные с помощью тематического моделирования темы преобразовываются во временны*е* ряды с помощью статистического *метода главных компонент* (Principal Components Analysis, PCA[7]). Главная задача преобразования данных – это их сжатие путем исключения избыточности, при минимизации потери информации.

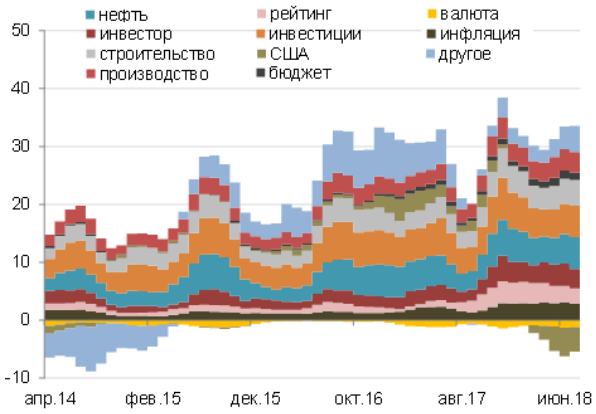
Регрессионный анализ показал, что статистически значимыми являются четыре главные компоненты, вклад в новостной индекс которых представлен ниже (Рисунок 6). Четыре главные компоненты объясняют 83% общей дисперсии индекса.

**Рисунок 6. Разложение новостного индекса на главные компоненты, пункты**




*Источники: интернет-новости, расчеты ДИП.*

**Рисунок 7. Разложение первой главной компоненты по темам, пункты**




*Источники: интернет-новости, расчеты ДИП.*

Если разложить по темам первую главную компоненту, которая по определению объясняет наибольшую долю вариации (до 40–50%), то можно заметить, что основными ее драйверами являются слова «нефть», «производство», «инвестиции», «валютный курс» (Рисунок 7). В целом общая динамика первой компоненты остается положительной, что оказывает благоприятное влияние на экономическую

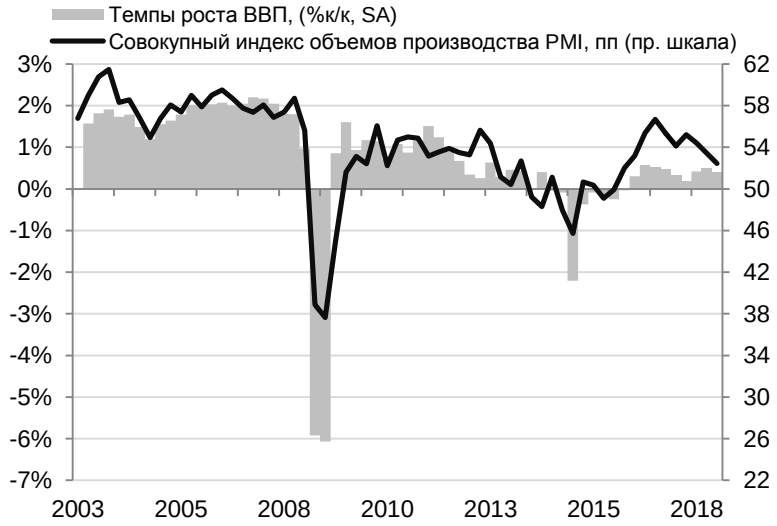
[7] Principal Component Analysis – один из методов факторного анализа, задача которого предоставить данные в меньшей размерности, минимизировав потерю информации. Его алгоритм заключается в переходе от изначальных данных к нескольким новым группам, в которых данные имеют схожие взаимосвязи.

активность. Иными словами, чем чаще встречаются в новостях упомянутые выше слова, являющиеся драйверами почти половины новостного индекса, тем выше при прочих равных условиях оказывается наша оценка деловой активности на базе экономических новостей.

*Насколько же текущий новостной индекс полезен в анализе и прогнозировании краткосрочных изменений деловой активности в России?*

Чтобы ответить на этот вопрос, мы использовали все 50 тем, выявленные с помощью модели LDA, и попытались соотнести этот информационный массив с наблюдаемым «эталонным» состоянием деловой активности. В качестве последнего нами был использован композитный индекс деловой активности PMI[8]. Выбор индекса PMI как критерия качества новостного индекса был обусловлен его сильной корреляцией с основным экономическим показателем – валовым внутренним продуктом (ВВП) (Рисунок 8). ВВП не был использован в качестве прогнозного показателя, поскольку статистика по нему публикуется на ежеквартальной основе, то есть редко и со значительным запаздыванием во времени. Мы использовали в качестве контрольной точки период с февраля 2017 г. по сентябрь 2018 гг., а остальные месяцы – в качестве периода обучающих наблюдений (апрель 2014 – январь 2018 г.).

**Рисунок 8. ВВП и композитный индекс PMI в России**

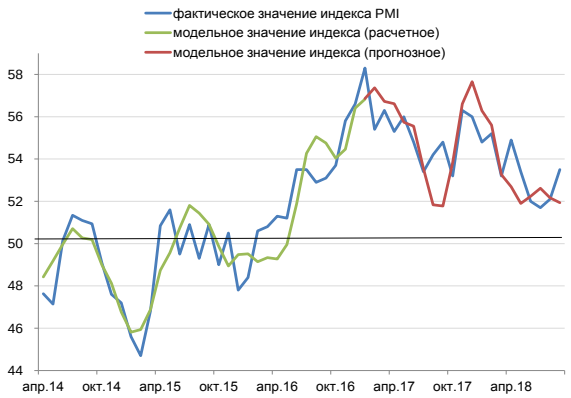



*Источники: Росстат, IHS Markit.*

Результаты, иллюстрирующие качество попадания в композитный PMI-индекс с помощью построенного новостного индекса, графически проиллюстрированы ниже (Рисунок 9, Рисунок 10).

[8] Композитный индекс PMI – это средневзвешенное значение индекса объема производства обрабатывающих отраслей и индекса деловой активности в сфере услуг. Индекс PMI обрабатывающих отраслей основан на пяти ключевых показателях со следующим удельным весом: новые заказы – 0,3; объемы производства – 0,25; занятость – 0,2; сроки поставок сырья и материалов – 0,15; запасы сырья и материалов – 0,1. Индекс PMI сектора услуг рассчитывается путем взвешивания процентного соотношения ответов респондентов со следующим весом: «улучшение/рост» – 1,0; «неизменность» – 0,5; «ухудшение/снижение» – 0,0.

**Рисунок 9. Индекс PMI и новостной индекс, пункты**




*Источники: IHS Markit, интернет-новости, расчеты ДИП.*

**Рисунок 10. Новостной индекс и индекс деловой активности PMI, пункты (скользящее среднее за три месяца)**


*Источники: IHS Markit, интернет-новости, расчеты ДИП.*

Прогноз индекса PMI показал, что модель, построенная на основе текстового анализа, имеет удовлетворительную прогнозную силу. Можно увидеть, что обозначенные нами темы зачастую «улавливают» общие тренды, но в отдельные периоды могут и не отражать некоторых поворотных моментов в экономической динамике или отражать их недостаточно корректно.

Следует отметить, что новостной индекс по построению предполагает более стабильную динамику по сравнению с фактическим изменением PMI-индекса: балансовые оценки из опросов компаний от месяца к месяцу могут изменяться потенциально сильнее, чем риторика текстов экономических новостей, которая отличается большей устойчивостью. Как следствие, мы допускаем временные существенные расхождения в динамике новостного индекса и композитного PMI-индекса на горизонте одного-двух месяцев. Однако на чуть более длинных временных горизонтах новостной индекс продолжает стабильно «улавливать» тренды в изменении общих оценок состояния деловой активности, которые следуют из данных опросов.

Если же обращаться непосредственно к оценке текущего уровня деловой активности на основе данных экономических новостей, то следует отметить, что построенный нами новостной индекс находится вблизи отметки 52 пункта с мая 2018 года. Это говорит о сохранении стабильных, но умеренных темпов роста экономики, указывая на риски их небольшого замедления во втором полугодии 2018 года. При этом по большинству новостных тем произошло снижение позитивной динамики. Наши оценки показывают, что количество положительных новостей снизилось прежде всего в темах, связанных с инфляцией, долговым рынком, банковским сектором. Однако данное снижение было в значительной степени компенсировано снижением волатильности на финансовых рынках. Последнее наблюдение подтверждает тот факт, что нормализация ситуации на финансовых рынках выступает в качестве важного стабилизирующего фактора и является индикатором улучшения ситуации в экономике. На аналогичный вывод Департамент исследований и прогнозирования (ДИП) указывал в мае 2016 г., основываясь на проведенных совместно с

Институтом экономической политики опросах предприятий промышленности и сельского хозяйства[9].

Дополнительно следует отметить, что понижательное давление на индекс периодически оказывает тема, связанная с США и геополитикой. Однако важно подчеркнуть, что влияние последней темы на новостной индекс, согласно нашим оценкам, носит кратковременный характер и не оказывает устойчивого негативного влияния на наши оперативные оценки состояния деловой активности российской экономики.

Достойны внимания и еще несколько прогностических свойств новостного индекса, которые нам удалось обнаружить. Так, наивысшую прогнозную силу индекс показывает во вторую неделю месяца (Рисунок 12). Это может быть объяснено тем, что менеджеров (на основе ответов которых построен индекс деловой активности PMI), как правило, начинают опрашивать с конца второй – начала третьей недели календарного месяца. Поэтому в большей степени на мнение участвующих в опросе менеджеров влияют события, произошедшие в середине месяца.

**Рисунок 11. Индекс, построенный по данным первой недели каждого месяца, пункты**

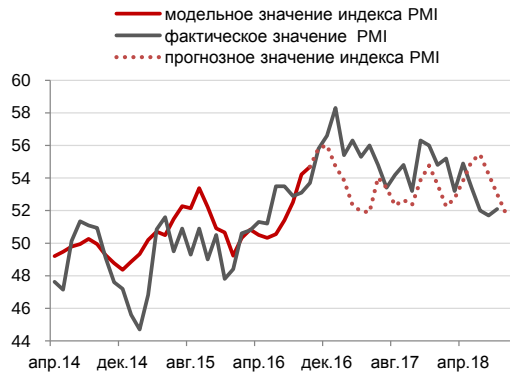



**Рисунок 12. Индекс, построенный по данным второй недели каждого месяца, пункты**

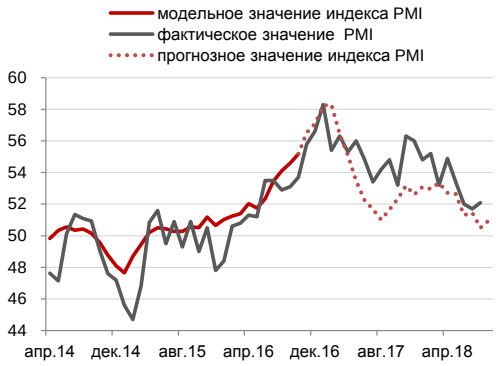



**Рисунок 13. Индекс, построенный по данным третьей недели каждого месяца, пункты**

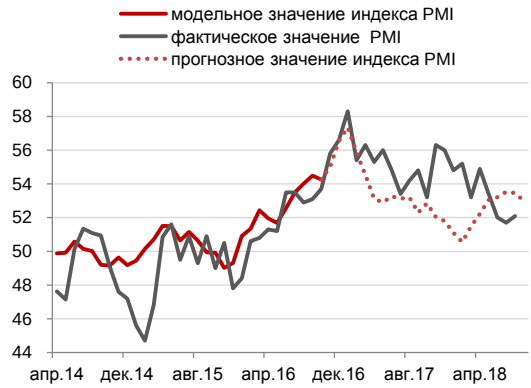
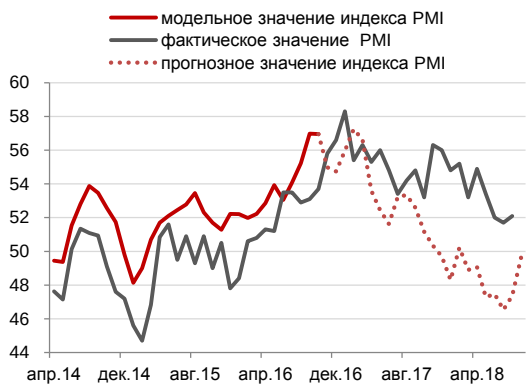



**Рисунок 14. Индекс, построенный по данным четвертой недели каждого месяца, пункты**

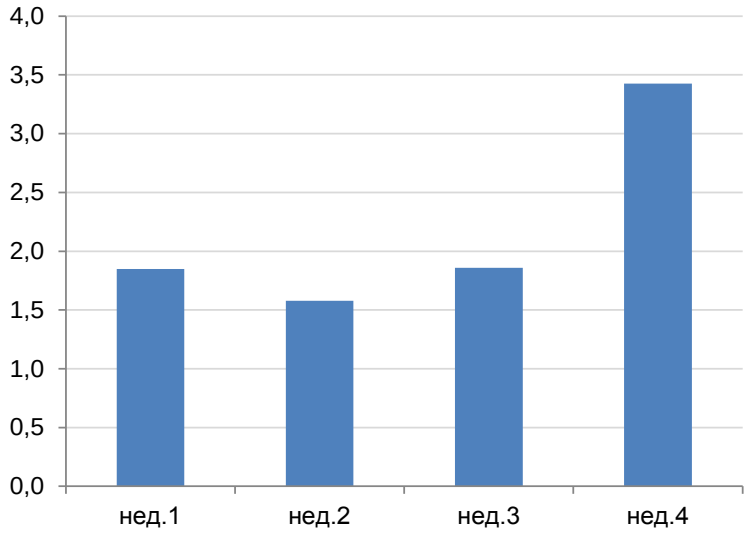



*Источники: интернет-новости, расчеты ДИП.*

[9] См. Бюллетень «О чем говорят тренды» (№ 6, май 2016 г.). В фокусе. Предприятиям нужен разный курс рубля: результаты опроса.

В среднем же качество прогноза оценок деловой активности в заданный месяц не имеет выраженной тенденции к улучшению по мере накапливания новостей в течение каждой последующей недели месяца (Рисунок 15). Массив экономических новостей за полный месяц, который покрывает период в одну-две недели после опросов менеджеров, показывает заметно более низкое качество прогноза PMI-индекса.

**Рисунок 15. Средние абсолютные ошибки индекса по неделям, пункты**




*Источники: интернет-новости, расчеты ДИП.*

В качестве потенциального объяснения этому может послужить несколько причин. Выделим три из них, которые, с нашей точки зрения, являются основными:

1. *Периодичность опросов менеджеров*. События на четвертой неделе календарного месяца происходят уже после опроса менеджеров, и они с точки зрения данных PMI-индексов будут влиять на оценки респондентов в отношении состояния деловой активности уже не ранее следующего календарного месяца.
2. *Выход краткосрочной статистики Росстата во второй половине месяца*. Комментарии в экономических новостях публикуются в течение нескольких дней после выхода ежемесячных данных Росстата по состоянию экономической активности. Согласно официальному графику Росстата, данные всегда публикуются во второй половине или даже ближе к концу месяца. При этом краткосрочная макроэкономическая статистика в силу целого ряда причин (в том числе методологического характера) традиционно характеризуется высокой волатильностью: на высокочастотных данных зачастую бывает проблематично найти важные тренды и/или переломные точки в динамике экономической активности. Так, в экономической прессе могут встречаться комментарии с неоправданно негативной тональностью применительно к краткосрочным данным. Например, отрицательная тональность может присваиваться снижению в заданном меся-

це краткосрочного показателя в терминах *«год к году»* (в то время как это может объясняться главным образом эффектом высокой базы прошедшего года) или *«месяц к месяцу»* без поправки на сезонный фактор (когда в действительности это снижение вписывается в привычную сезонность краткосрочного показателя экономической активности). В целом же выход краткосрочной информации Росстата происходит уже после опроса менеджеров и должен отражаться на их решениях не ранее следующего опросного месяца, если событие оказалось очень значимым.

3. *Особенности построения PMI-индекса*. В фокусе экономической прессы в конце месяца часто оказывается ограниченная группа показателей деловой активности. Например, ежемесячная краткосрочная статистика Росстата по производству часто комментируется в экономических СМИ с акцентом на динамике выпуска продукции в отраслях обрабатывающего сектора, в то время как обработке присваивается меньший вес по сравнению со сферой услуг в опросах, на основе которых строятся PMI-индексы.

Как результат, поведение новостного индекса, построенного на информационном массиве с включением последней недели месяца, может идти вразрез с динамикой PMI-индекса. Тем не менее описанные выше свойства новостного индекса, с нашей точки зрения, означают, что критерий качества построенного аналитического показателя ни в коем случае не должен замыкаться исключительно на ошибке прогнозирования показателей официальной статистики и опросных показателей, в особенности на краткосрочных горизонтах. Важно, что новостной индекс, как мы уже отмечали, «улавливает» общие тренды в динамике экономической активности, но при этом может трактоваться и как полноценный независимый инструмент оценки деловой активности, который является возможной альтернативой и дополнением к существующим статистическим и опросным показателям. Более детальное изучение данного вопроса, по нашему мнению, должно лечь в основу дальнейшей исследовательской работы по развитию индикаторов состояния экономики, рассчитываемых с помощью текстового анализа.

***

В ходе проведенного анализа мы показали, что новости сами по себе способны достаточно качественно описывать произошедшие в экономике события. Безусловно, в новостях существуют темы, входящие в индекс, которые мы не всегда сможем точно интерпретировать. Это связано как с человеческим фактором, так и с особенностями машинных методов. Несмотря на объективные преимущества машинного анализа текстов, он также имеет свои слабые стороны, которые связаны главным образом с проблемами определения тематики и тональности текстового документа, а также предварительной обработки текста – в особенности русскоязычных слов, написание которых подчинено многим правилам и исключениям.

Объективные методологические сложности, которые сопровождают работу с неструктурированными большими массивами данных, а также продолжающееся активное развитие подходов к текстовому анализу, безусловно, оставляют значительное пространство для доработки и совершенствования подходов к работе с текстами при анализе и прогнозировании экономики. С точки зрения практических выводов для экономистов они пока все же указывают на необходимость взвешенно и осторожно интерпретировать получаемые с помощью текстового анализа результаты.

Предложенный нами новостной индекс уже в первом приближении дает удовлетворительные оперативные оценки состояния деловой активности в российской экономике, которая может активно применяться в том числе в практике центрального банка. Но действительно ли способны методы текстового анализа привносить дополнительную полезную составляющую в информацию о состоянии экономики, которая содержится в официальной статистике? На сегодняшний день все же трудно дать однозначный ответ на данный вопрос, к нему нас может приблизить только дальнейшее развитие методов обработки неструктурированных массивов данных. Поэтому на текущем этапе построенный нами новостной индекс правильнее будет воспринимать как информацию к размышлению, отправную точку в широком комплексе экономических исследований, посвященных анализу больших объемов неструктурированной информации, которые в дальнейшем будут проводиться в Банке России.

## ЛИТЕРАТУРА


1. Воронцов К. (2013) Вероятностное тематическое моделирование. URL: http://pratsi.opu.ua/app/webroot/articles/1414145257.pdf (дата обращения – 01.10.2018)
2. Яковлева К. (2017). Оценка экономической активности на основе текстового анализа // Серия докладов об экономических исследованиях в Банке России, № 25
3. Ardia D., Bluteau K. (2017). Questioning the News About Economic Growth: Sparse Forecasting Using Thousands of News-Based Sentiment Values. Preprint submittes to SSRN, July 21
4. Bholat D., Hansen S. (2015). Text mining for central banks. Centre for Central Banking Studies
5. Blei D., Ng A., Jordan M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research
6. Bloom N., Baker S., Davis S. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics
7. Doms M., Morin N. (2004). Consumer Sentiment, the Economy, and the News Media. Finance and Economics Discussion Series (FEDS)
8. Hendry S, Madeley A. (2010). Text Mining and the Information Content of Bank of Canada Communications. Staff Working Paper 2010–31
9. Kholodilin K., Thomas T., Ulbricht D. (2017). Do media data help to predict German industrial production? Journal of Forecasting

10. Nyman R, Ormerod P. (2014). Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis
11. Shapiro A., Sudhoh M., Wilson D. (2017). Measuring News Sentiment. Federal Reserve Bank of San Francisco Working Paper Series
12. Thorsrud A. (2016). Words are the new numbers: A newsy coincident index of business cycles. Norges Bank Research. Working Paper
13. Turrell A., Speigner B. (2018). Using job vacancies to understand the effects of labour maket mis-match on UK output and productivity. Staff Working Paper. № 737

**Департамент исследований и прогнозирования**

Ксения Яковлева
Алексей Поршаков