

ПОСТРОЕНИЕ ИНДЕКСА ПРЕДПРИНИМАТЕЛЬСКОЙ УВЕРЕННОСТИ В РОССИИ НА ОСНОВЕ АНАЛИЗА ТОНАЛЬНОСТИ НОВОСТНЫХ ТЕКСТОВ В ИНТЕРНЕТЕ

Ф. В. УЛЬЯНКИН

А. В. ПОЛБИН

В работе, на основе новостных текстов, строится индекс предпринимательской уверенности для России. В качестве исходных данных используются новостные статьи, взятые из интернет-ресурсов. Обработка новостных статей производится методами текстового анализа. Сентимент-окрас статей оценивается с помощью словаря, собранного на основе краудсорсинга. Анализ динамики индекса показывает, что экономические агенты негативно оценивают меры поддержки бизнеса, связанные с коронакризисом. После пика в марте 2020 г., связанного с анонсированием первых мер помощи бизнесу, индекс претерпевает самый продолжительный спад за последние три года. Анонсирование новых мер поддержки не возвращает индекс к прежним показателям.

Ключевые слова: текстовый анализ, анализ тональности, индекс предпринимательской уверенности, машинное обучение.

Введение

Для заблаговременного выявления экономической нестабильности обычно используются системы опережающих индикаторов – барометров общеэкономической конъюнктуры. Они помогают заниматься мониторингом и прогнозированием деловой активности, а также сокращать временные интервалы, необходимые для принятия упреждающих решений, важных для стабилизационной политики государства.

Поначалу подобные индикаторы строились вручную в ходе социологических опросов. Например, долгое время Сбербанк совместно с Левада-центром строил с помощью опросов в разных регионах индекс финансовых настроений¹.

Начиная с работы [4] пришло осознание, что построение барометров для разных рынков можно автоматизировать. В ходе взаимо-

действия человека с Интернетом вырабатывается огромное количество неструктурированной информации, которую можно было бы использовать для понимания того, как экономические агенты оценивают текущее состояние дел.

Сегодня весь мир переживает кризис, вызванный пандемией коронавируса. Правительства принимают экономические меры, призванные поддержать бизнес, несущий потери из-за режима самоизоляции. Многие эксперты заявляют, что меры поддержки российской экономики недостаточны². В связи с этим становится актуальной проблема конструирования индикатора, отражающего уровень неопределенности, связанный с ведением бизнеса и характеризующий предпринимательскую уверенность. В настоящей работе на основе потока новостей мы делаем попытку построить такой индикатор.

Ульянкин Филипп Валерьевич, аспирант РАНХиГС при Президенте Российской Федерации (Москва), e-mail: filfonul@gmail.com; Полбин Андрей Владимирович, заведующий лабораторией математического моделирования экономических процессов РАНХиГС при Президенте Российской Федерации; заместитель заведующего международной лабораторией математического моделирования экономических процессов Института экономической политики имени Е.Т. Гайдара, канд. экон. наук (Москва), e-mail: apolbin@iep.ru

¹ URL: <https://www.sberbank.ru/ru/about/analytics/mood>

² Меры поддержки экономики недостаточны, и власти это понимают // Reuters. 9 апреля 2020 г. URL: <https://ru.reuters.com/article/businessNews/idRUKCN21R1BJ-ORUBS>; Слепое пятно в глазах государства // Ведомости. 5 апреля 2020 г. URL: <https://www.vedomosti.ru/opinion/columns/2020/04/05/827199-slepoe-pyatno>; Титов счел недостаточными меры поддержки бизнеса в условиях вируса // РБК. 30 марта 2020 г. URL: <https://www.rbc.ru/economics/30/03/2020/5e8230d89a7947e55652c618>

Обзор литературы

Для автоматического построения барометров финансовой конъюнктуры обычно используют либо поисковые запросы, либо комментарии в социальных сетях, либо поток новостей из СМИ. Частотная методология предполагает, что индексы строят, отталкиваясь от вручную отобранных слов (дескрипторов). Для каждого дня подсчитывается, как часто встречаются дескрипторы, после чего их взвешивают на подобранные по выбранной автором методологии веса. Итоговый ряд сглаживается.

Одними из первых барометры финансовой конъюнктуры стали строить Вэриан и Чои. В статьях [4] и [5] они показали, что индексы, основанные на поисковых запросах, улучшают прогноз безработицы и спроса на автомобили.

Для России, по аналогичной методологии, один из первых барометров финансовой конъюнктуры был построен в работе Столбова [2]. Он проанализировал, какие поисковые запросы по категории «Финансы и страхование» пользователи Google делали в пик кризиса 2008 г. Динамику поисковых запросов, взвешенную на взаимные корреляции, автор использовал в качестве индикатора финансовой нестабильности. В статье было показано, что такой индикатор улучшает прогноз динамики депозитов физических лиц.

Бэйкер, Блум и Дэвис в работе [3] разработали индекс экономической неопределенности. Они проводили расчет, как часто в газетах встречаются выбранные ими кризисные слова. В конечном счете авторы обнаружили, что их индекс демонстрирует интерпретируемую динамику, принимает высокие значения в окрестности серьезных дебатов относительно фискальной политики, а также в окрестности других событий, связанных с неопределенностью будущей экономической политики. Такие вычисления были проведены для 11 разных стран, в том числе и для России на основе га-

зеты «Коммерсантъ». Все построенные индексы с подробным описанием методологии авторы публикуют на регулярной основе на сайте своего проекта³.

Голощапова и Андреев [1] попытались на основе сентимент-окраса пользовательских комментариев в СМИ построить индекс неопределенности инфляционных ожиданий, а также индекс, описывающий их интенсивность. Индикаторы получились релевантными основным макроэкономическим трендам. На сайте проекта⁴, кроме индекса инфляционных ожиданий, авторы также строят индекс финансовой неопределенности, подсчитывая, по аналогии с подходом Блума и соавторов, как часто в СМИ встречается ряд кризисных дескрипторов.

В настоящем исследовании разрабатывается собственная методология для построения индекса предпринимательской уверенности. Мы работаем с новостными текстами из нескольких крупных СМИ. На основе нескольких правил мы выделяем поток новостей, связанных с бизнесом в России. Вслед за чем устанавливаем сентимент-окрас каждой статьи и анализируем его динамику.

Данные и методология

Для построения индекса мы собрали корпус новостных текстов на основе публикаций крупнейших новостных ведомств в социальных сетях. В выборку вошли «Интерфакс», «Коммерсантъ», «Ведомости», «Лента», «РБК», «Медуза», «Российская газета», «РИА-новости», «РТ», «ТАСС». Всего в выборку попало более миллиона текстов. Их подавляющее большинство было создано в период с 1 января 2017 г. (более 700 тыс. статей). В год выпускалось примерно 200 тыс. статей.

Корпус текстов был очищен от знаков пунктуации, приведен к нижнему регистру, все буквы «ё» были заменены на «е». Для большей эффективности по скорости работы удаление

³ URL: <http://www.policyuncertainty.com/index.html>

⁴ URL: <http://bigdata-indicators.com/index.html>

знаков пунктуации осуществлялось с помощью регулярных выражений. Далее тексты были очищены от стоп-слов, токенизированы и нормализованы посредством лемматизации.

Вслед за тем названия статей и их краткие описания, вынесенные в социальную сеть (сниппеты), были обработаны с помощью тонального словаря русского языка⁵. Словарь собирался проектом «Карта слов» посредством краудсорсинга. В процессе разметки отвечающему предлагалось оценивать разные слова как нейтральные, положительные или отрицательные; также был предусмотрен ответ «не знаю». По разметке были рассчитаны показатели, отвечающие за уверенность в вердикте. Все вердикты, обладающие высокой степенью рассогласованности, удалялись. На практике высокая рассогласованность возникает в ситуациях, когда оценка слова сильно зависит от контекста. Дополнительно были удалены все слова, связанные с медициной, вирусами и эпидемией.

Из потока новостей отфильтровывались только те, которые содержали токены «малый_ADJ», «бизнес_NOUN», «экономика_NOUN». Дополнительно шла только фильтрация новостей, относящихся к России. После фильтрации в рассмотрении осталось 13 145

статей. Медиана числа статей за день оказалась равна 10.

Все слова в названии и сниппете статьи заменялись на их «силу выраженности эмоционально-оценочного заряда», полученную по разметке. Положительные числа означали, что слово обладает позитивной коннотацией. Отрицательные числа говорили о негативной коннотации. Чем сильнее была выражена эмоция, тем больше по модулю была оценка слова. Если слова не оказывалось в словаре, оно заменялось на ноль.

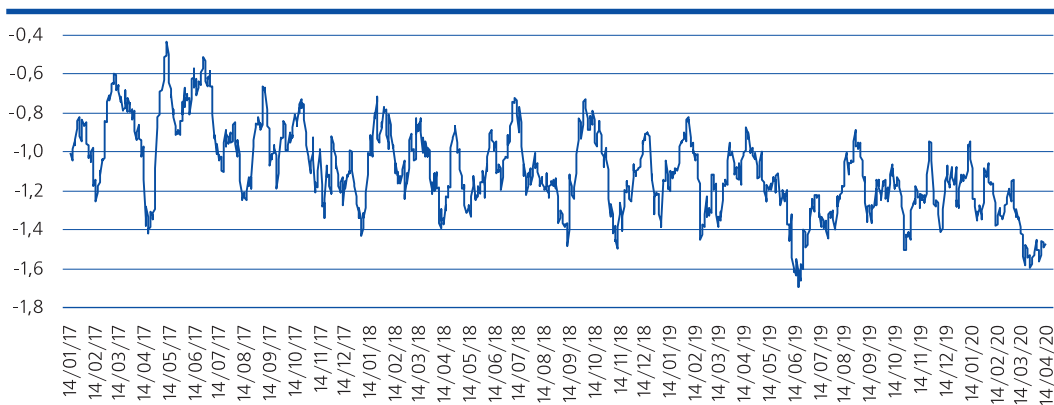
Итоговая оценка коннотации для статьи находилась как сумма оценок слов. В рамках одного дня считалась средняя оценка по всем статьям. Далее, для того чтобы акцентировать внимание на долгосрочных изменениях, а не на междневных колебаниях, ряд сглаживался 14-дневным скользящим средним.

В итоге был получен индекс, описывающий сентимент-окрас новостного фона, связанного с российской бизнес-активностью. Чем ниже индекс, тем больше негативной информации, в среднем, содержится в потоке.

Динамика индикатора

Индикатор охватывает промежуток с 1 января 2017 г. по 1 мая 2020 г.. Динамика индекса за

Рис. 1. Динамика индекса с 1 января 2017 г. по 1 мая 2020 г.



Источник: расчеты авторов.

⁵ URL: https://github.com/dkulagin/kartaslov/tree/master/dataset/emo_dict

весь рассматриваемый период отображена на рис. 1, за последний год – на рис. 2.

Как показано на рисунках, с конца марта текущего года по настоящее время произошло резкое падение индекса. В течение месяца он находится на низком уровне. Все предыдущие его падения были кратковременными и длились не более недели. (Последнее, самое глубокое сопоставимое падение, наблюдалось в непродолжительный период в 20-х числах июля 2019 г.)

Рассмотрим подробнее новости с негативной коннотацией, которые попадают в три самых длительных «провала», произошедших в течение последнего года. В 20-х числах июля 2019 г. одновременно наблюдались нескольких крупных скандалов, связанных с развитием бизнеса в России: уголовное дело против «Рольфа»; разрыв отношений с Грузией и проблемы с рынком вина; проблемы с «мусорной» реформой и операторами по его уборке; скандал вокруг расследования Голунова о том, кому принадлежит бизнес ритуальных услуг. Эти события, случившись одновременно, создали кратковременный рост неопределенности, который зафиксировал наш индекс в виде снижения предпринимательской уверенности.

Примеры новостных заголовков за период с 26 июня по 4 июля 2019 г.:

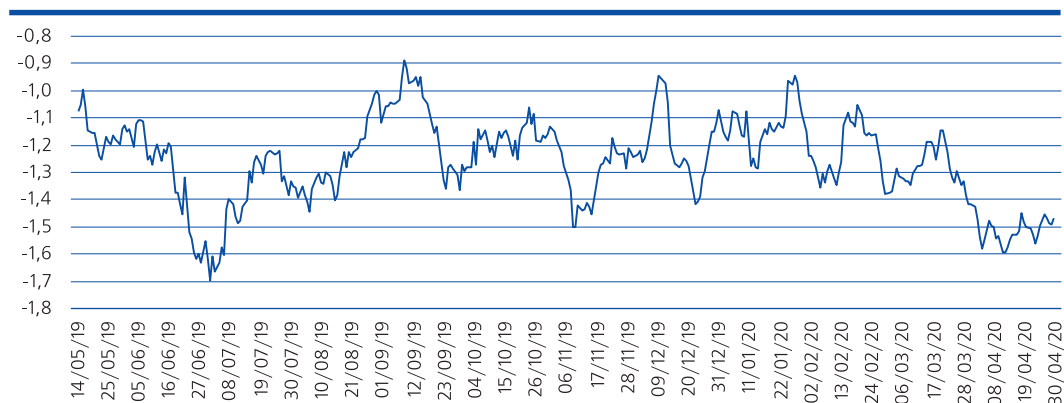
- Кредиты – не главная причина падения реальных располагаемых доходов, считает ЦБ РФ.
- Как люди будут платить за вывоз мусора.
- Задержан первый подозреваемый по делу «Рольфа». Обыски прошли во всех дилерских центрах компании.
- Налоговая служба заблокировала счета бизнес-школы и крупнейшего частного университета «Синергия».
- Бизнес-омбудсмен ищет управу на силовиков в Совете безопасности.
- Двое полицейских начальников устроили торговлю наркотой прямо у себя в отделе.
- Грузия лишилась российского рынка вина.
- Бизнес могут обязать оплачивать инвест-программу «Россетей».
- Кто владеет московскими кладбищами?

В период с 9 по 14 ноября 2019 в СМИ появилось довольно большое число статей на тему «5 лет назад в России начался экономический кризис». (Кратковременное «дно» на рисунках соответствует именно этим статьям.)

Примеры новостных заголовков за период с 9 по 14 ноября 2019 г.:

- Пять лет назад в России начался экономи-

Рис. 2. Динамика индекса с 1 мая 2019 г. по 1 мая 2020 г.



Источник: расчеты авторов.

ческий кризис. Страна от него так и не оправилась.

- Путин посетовал на недобросовестную конкуренцию в мировой экономике.
- Как госорганы проверяли малый бизнес в 2019 г.
- Путин рассказал, что удержало экономику от сползания к рецессии.
- Крупный в России автодилер «Рольф» выставлен на продажу.
- Холдинг Вексельберга просит избавить аэропорты от бессмысленных требований по безопасности.
- Билеты среднего ценового сегмента сейчас продаются гораздо хуже, чем несколько лет назад.

В период с 25 марта по 1 мая 2020 г. встречалось огромное число статей о том, что в нынешних реалиях малый бизнес чувствует себя чрезвычайно плохо и пессимизм нарастает.

Примеры новостных заголовков за период с 25 марта по 1 мая 2020 г.:

- Пессимизм в бизнес-среде достиг уровня кризиса 2015 г.
- Банки заявили о росте активности мошенников в период самоизоляции.
- Миллиардер Агаларов сравнил свой бизнес с тонущим кораблем.
- АСИ предложило Мишустину ряд мер по поддержке занятости и доходов россиян.
- Малому бизнесу не хватает до зарплаты.
- Бизнес-омбудсмен Титов пожаловался Путину на отсутствие адекватного ответа государства на кризис.
- Правительству неплохо бы переселиться на «планету людей».
- В Госдуме пожаловались на реализацию мер поддержки малого и среднего бизнеса.

Последний положительный пик в индексе перед текущим резким спадом соответствовал 20–22 марта. В этот период в СМИ появлялся пласт новостей, в которых описывалось, что правительство обсуждает меры поддержки бизнеса в условиях коронавируса.

Примеры новостных заголовков за 20–22 марта 2020 г.:

- Кабмин обсудил меры поддержки граждан и малого бизнеса.
- Мишустин рассказал о влиянии коронавируса на экономику.
- Минфин обнулil таможенную пошлину на импорт лекарств.
- Правительство подготовило новый план поддержки экономики из-за вируса.
- Глава ЦБ Эльвира Набиуллина рассказала о ситуации в экономике России.
- Мишустин распорядился дать бизнесу отсрочку по арендным платежам.
- Чем грозит россиянам борьба с эпидемией.
- Как ритейлеру защитить бизнес от коронавируса?

Отдельно стоит отметить, что, несмотря на все меры, предпринятые для фильтрации новостей, в выборку попали нерелевантные статьи. Например, статьей с самым положительным окрасом за рассматриваемый период оказалась новость об успешном запуске ракеты с космодрома «Восточный». К сожалению, из-за семантического разнообразия невозможно отсеять все нерелевантные статьи. Тем не менее при просмотре случайной выборки новостных заголовков за разные даты существенного смещения замечено не было — в связи с этим нерелевантные статьи не должны существенно исказить динамику индикатора.

Выводы

В настоящей работе предложен индикатор предпринимательской уверенности, показывающий динамику негативной информации в статьях, связанных с российским бизнесом. Такая информация определяется коннотацией конкретных слов, вошедших в статью. Анализ событий, соответствующих самым продолжительным спадам, сигнализирует о том, что индекс в целом отражает основные негативные события, связанные с бизнесом в России.

В течение последних месяцев индекс претерпевает самый продолжительный спад за

последние три года. В СМИ работу бизнеса сопровождает стабильный негативный информационный фон. После принятия первого пакета мер по поддержке бизнеса в нем наблюдался резкий всплеск — экономические агенты отнеслись к нему позитивно. Вслед за тем индекс пошел на спад и не возвращался к таким

значениям ни после принятия новых мер, ни после выступлений президента. Это позволяет нам говорить не только о том, что российский бизнес переживает период серьезного кризиса, но и о том, что экономические агенты негативно воспринимают государственные меры по его поддержке. ■

Литература

1. Голощапова И.О., Андреев М.Л. Оценка инфляционных ожиданий российского населения методами машинного обучения // Вопросы экономики. 2017. № 6. С. 71–93.
2. Столбов М. Статистика поиска в Google как индикатор финансовой конъюнктуры // Вопросы экономики. 2011. № 11. С. 79–93.
3. Baker S. R., Bloom N., Davis S. J. Measuring Economic Policy Uncertainty // Quarterly Journal of Economics. 2016. Vol. 131. No. 4. Pp. 1593–1636.
4. Choi H., & Varian H. Predicting initial claims for unemployment benefits // Google Technical Report. 2009. Pp. 1–5.
5. Choi H., Varian H. Predicting the present with Google Trends // Economic Record. 2012. Vol. 88. Pp. 2–9.

References

1. Goloshchapova I., Andreev M. Measuring inflation expectations of the Russian population with the help of machine learning // Voprosy Ekonomiki. 2017. No. 6. Pp. 71–93.
2. Stolbov M. Statistics of Search Queries in Google as an Indicator of Financial Conditions // Voprosy Ekonomiki. 2011. No. 11. Pp. 79–93.
3. Baker S. R., Bloom N., Davis S. J. Measuring Economic Policy Uncertainty // Quarterly Journal of Economics. 2016. Vol. 131. No. 4. Pp. 1593–1636.
4. Choi H., & Varian H. Predicting initial claims for unemployment benefits // Google Technical Report. 2009. Pp. 1–5.
5. Choi H., Varian H. Predicting the present with Google Trends // Economic Record. 2012. Vol. 88. Pp. 2–9.

Construction of a Business Confidence Index for Russia Based on the Sentiment Analysis of News Texts from the Internet

Filipp V. Ulyankin — PhD student at the Russian Presidential Academy of National Economy and Public Administration (Moscow, Russia). E-mail: filfonul@gmail.com

Andrey V. Polbin — Head of Mathematical Modelling of Economic Processes Department of the Russian Presidential Academy of National Economy and Public Administration; Deputy Head of Mathematical Modelling of Economic Processes International Department of the Gaidar Institute, Candidate of Economic Sciences (Moscow, Russia). E-mail: apolbin@iep.ru

In this paper, we construct a business confidence index for Russia based on news texts from the Internet. News articles are processed via text analysis methods. We estimate sentiment of text based on special dictionary, assembled by crowdsourcing. Analysis of the index dynamics shows that the Russian business negatively evaluates support measures associated with coronavirus crisis. After a peak in March 2020, when the first announcement of business assistance was made, the index is showing the longest decline in the past three years. Even a new support measures announcement has not returned the index to its previous values.

Key words: text analysis; sentiment analysis; business confidence index; machine learning.