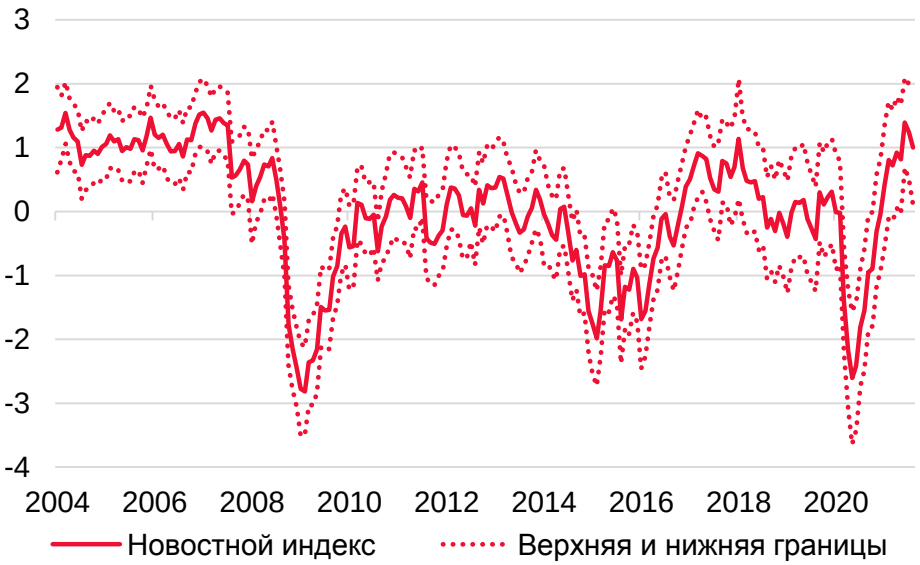
Оценка новостного индекса в августе 2021 года

- В августе[1] новостной индекс снизился до 1,0 п. с 1,2[2] п. по итогам июля, что может сигнализировать о некотором замедлении темпов роста экономической активности с повышенных уровней первого полугодия 2021 года. При этом значение индекса остается близким к максимальному на всей истории наблюдений (с 2004 года).
- Поддержанию индекса на повышенном уровне способствовали новости о высоком росте экономики во II квартале и повышении прогнозов роста экономики по итогам всего года. Позитивный вклад также внесли новости о расширении добычи нефти в рамках новых условий сделки ОПЕК+, остающихся повышенными объемах внешней торговли, продолжении восстановления сектора услуг.
- Основной вклад в снижение индекса, как и в июле, внесли новости об ухудшении эпидемический ситуации в стране и мире на фоне распространения дельта-штамма коронавируса, а также новости о продолжающихся глобальных перебоях с поставками, которые приводят к снижению производства товаров или сдерживают его рост на фоне расширения спроса.

Рисунок 1. Динамика новостного индекса



[1] Данные на 16 августа.
[2] Июльское значение индекса понизилось на 0,1 п.п. после добавления новостей за полный календарный месяц

О продолжении публикации ряда старого новостного индекса

Мы продолжаем расчет и публикацию старой версии новостного индекса в ближайшие месяцы (см. статистическое приложение к публикации) и просим читателей комментариев, а также пользователей данной статистики дать до конца сентября 2021 года обратную связь в свободной форме на почтовый ящик Департамента исследований и прогнозирования относительно необходимости продолжения публикации ряда новостного индекса по старой методологии, а также дать комментарии, замечания и предложения к новой версии.

Новая методология новостного индекса

Новостной индекс – это высокочастотный показатель, рассчитанный на основе ежедневных новостей для оценки экономической активности в стране. В предыдущем варианте (Оценка экономической активности на основе текстового анализа) он был откалиброван по динамике сводного индекса деловой активности PMI, который является одним из наиболее оперативных индикаторов изменения ситуации в экономики (публикуется гораздо раньше официальной экономической статистики). В качестве исходных данных использовались новостные статьи, которые обрабатывались методами текстового анализа и машинного обучения.

После падения в начале 2020 года старый новостной индекс остался на пониженном уровне, несмотря на то, что экономика начала активно восстанавливаться уже в начале III квартала 2020 года и к концу II квартала 2021 года вернулась на докоронавирусный уровень. Вместе с тем сонаправленность динамики новостного индекса и деловой активности в целом сохранялась. Такое изменение в динамике новостного индекса объясняется рядом причин, связанных с методологическими особенностями построения индекса и влиянием пандемии.

Ключевым фактором, по нашим оценкам, стало значительное сокращение числа доступных новостей. До настоящего времени при построении новостного индекса использовался всего один источник экономических новостей. Изначально выбор одного источника данных был обусловлен доступностью достаточно длинного временного ряда новостей, простотой веб-скрапинга[3] и легкостью расчета. Однако с июня 2020 г. число экономических новостей на указанном источнике сократилось в несколько раз, из-за чего экономическая ситуация предположительно стала освещаться заметно хуже, что в свою очередь негативно отразилось на качестве оценок новостного индекса.

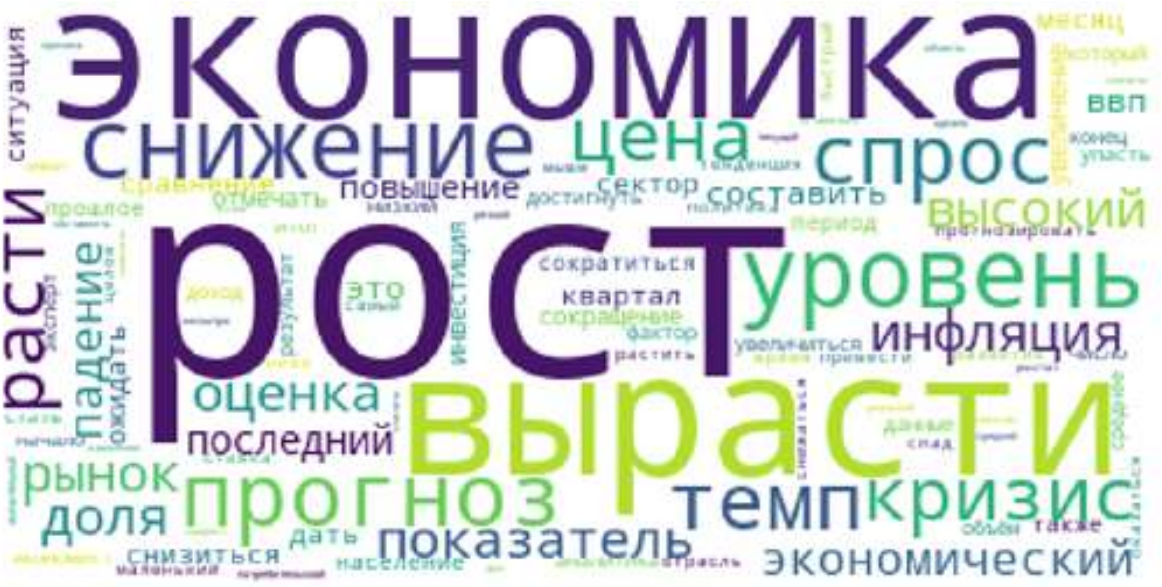
Важной причиной расхождения оценок новостного индекса и PMI, вероятно, стало появление в 2020 году новой темы – «коронавирус». Она имеет достаточно неоднозначную тональность и могла удерживать новостной индекс ниже отметки 50 п. даже в относительно благоприятные периоды. Так, в периоды восстановления экономической активности, зачастую сопровождавшиеся улучшением эпидемической ситуации и ослаблением ограничительных мер, тем не менее сохранялось большое количество заболевших, что могло оказывать негативное влияние на оценку новостного индекса.

[3] Технология, позволяющая автоматически сохранять текстовую информацию с интернет-сайтов.

Все это подтолкнуло нас к пересмотру методики построения новостного индекса. Набор новостей, используемых для оценки новостного индекса, значительно увеличен. Новая методология задействует сразу несколько специализированных источников деловых и экономических новостей со значительно большим количеством новостей, в том числе экономических, по сравнению с предыдущим источником. Также расширен исторический период, на котором рассчитывается новостной индекс (ранее – с 2014 года, в настоящий момент – с 2004 года). Теперь он включает период быстрого роста экономики начала 2000-х и мирового финансового кризиса 2008–2009 годов.

Мы постарались сделать новый новостной индекс максимально прозрачным с точки зрения методологии [4] и интерпретации. Новый индекс не калибруется по какому-либо макроэкономическому показателю. На первом этапе используется модель, автоматически выделяющая темы (100 тем). Из них по ключевым словам выбрана одна, максимально покрывающая экономическую тематику (см. Рисунок 2). Затем для каждой новости на основе обученной модели определяются ее релевантность экономической тематике (доля «экономической» темы в новости) и тональность на основе баланса слов с положительной и отрицательной смысловой окраской. Далее новости за определенный период времени (месяц) агрегируются в индивидуальные индексы по каждому источнику путем усреднения тональности новостей с учетом веса экономической тематики в них. Финальный индекс рассчитывается путем выделения общей тенденции в индивидуальных индексах методом главных компонент.

Рисунок 2. Ключевые слова для темы, соответствующей экономическим новостям (размер слова пропорционален важности слова в теме)



Мы справочно демонстрируем границы, которые сигнализируют о наличии значимого отклонения индекса от нуля. «Разлет» между этими границами зависит от неопределенности модели, которая находится в обратно пропорциональной зависимости от количества новостей (чем меньше новостей в текущем месяце доступно, тем выше неопределенность) и прямо пропорциональной – от разнонаправленности их тональности. Особенно

[4] Технические детали описаны в Приложении 1.

показательной является верхняя граница, снижение которой ниже нуля демонстрирует наличие значимого негативного новостного фона и совпадает с историческими моментами высокой турбулентности основных макроэкономических показателей. Технические детали новой методологии представлены в Приложении 1.

Хотя новый новостной индекс в целом не привязан при построении ни к какому экономическому показателю, он хорошо улавливает изменения в экономической динамике (Рисунки 3–5). Это указывает на возможность его использования для оперативного анализа текущей ситуации в экономике.

Важно отметить, что значение и динамика новостного индекса слабо зависят от выбора источника информации: графики новостных подындексов, посчитанных отдельно по каждому из новостных ресурсов, практически полностью совпали. Это доказывает, что значение новостного индекса определяется популярностью тех или иных тем, а не особенностями комментариев на эти темы.

Рисунок 3. Динамика нового новостного индекса и прироста ВВП (SA к/к, %)

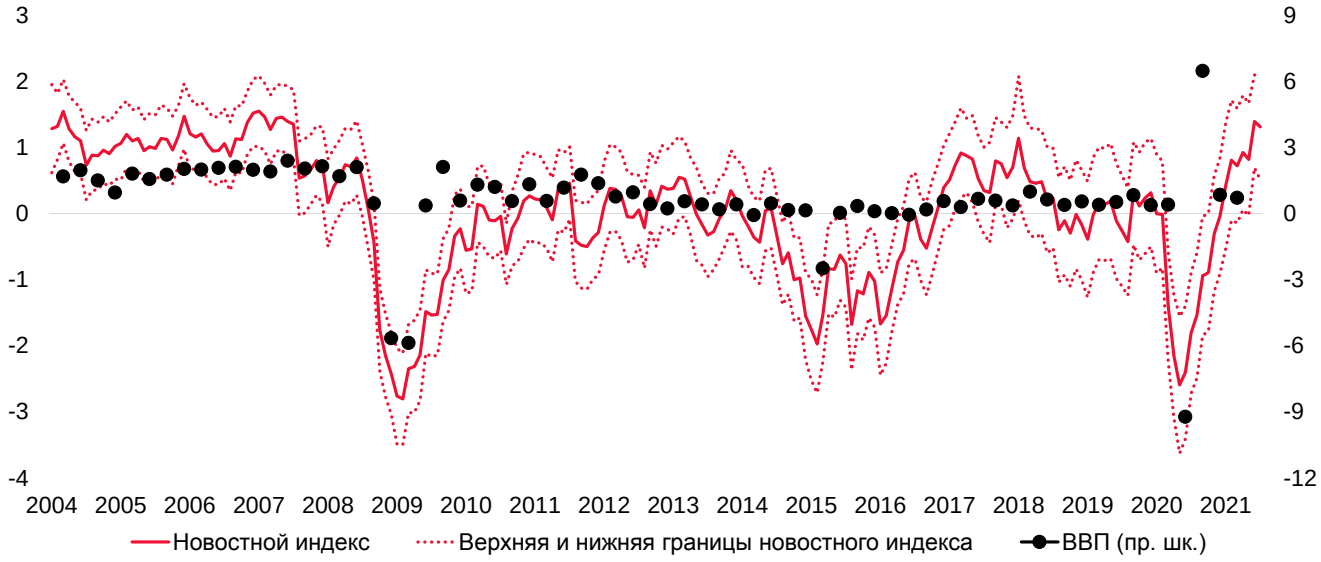


Рисунок 3а. Динамика нового новостного индекса и прироста ВВП (SA к/к, %)

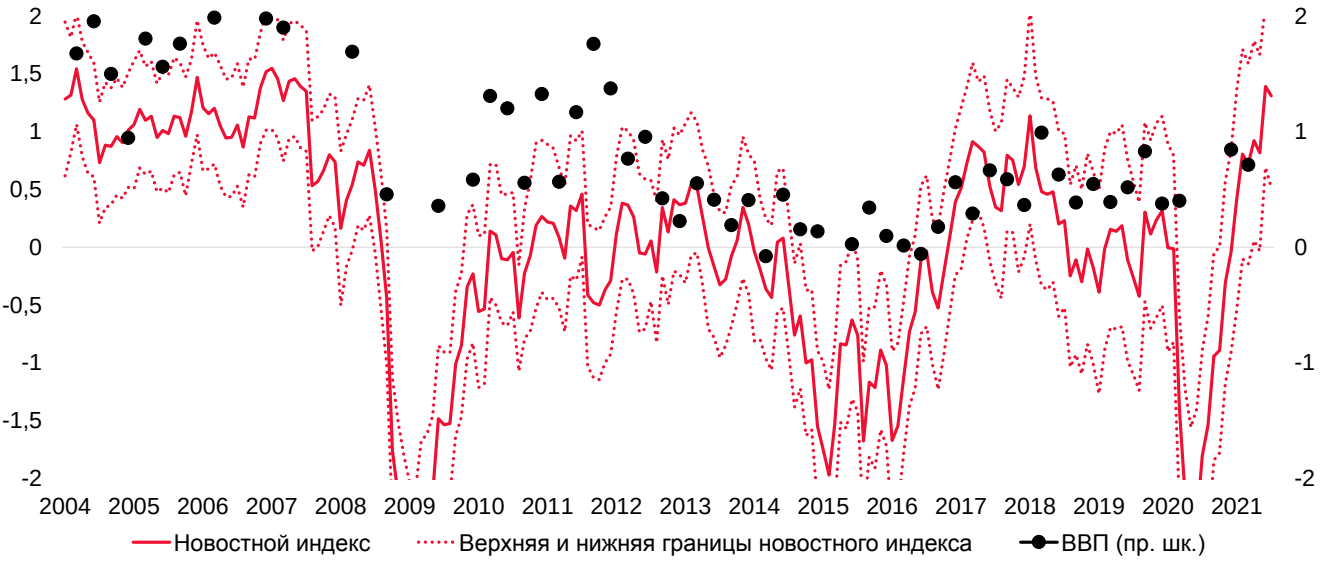


Рисунок 4. Динамика нового новостного индекса и инфляции (SA м/м, %)

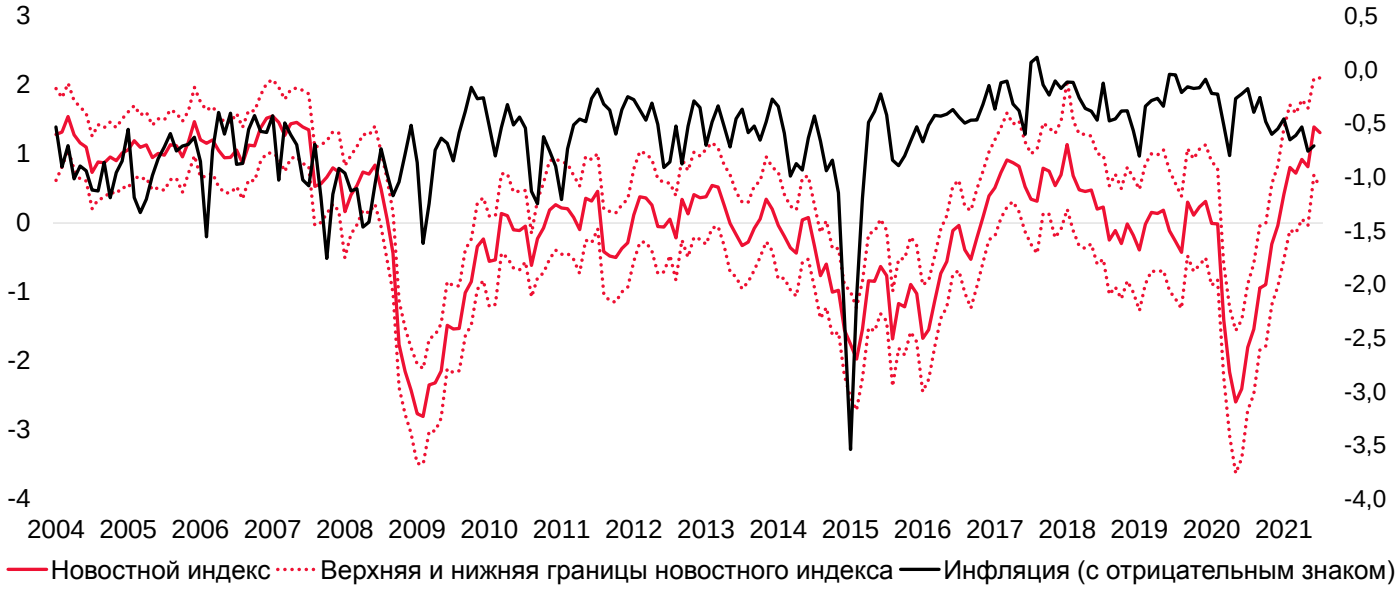
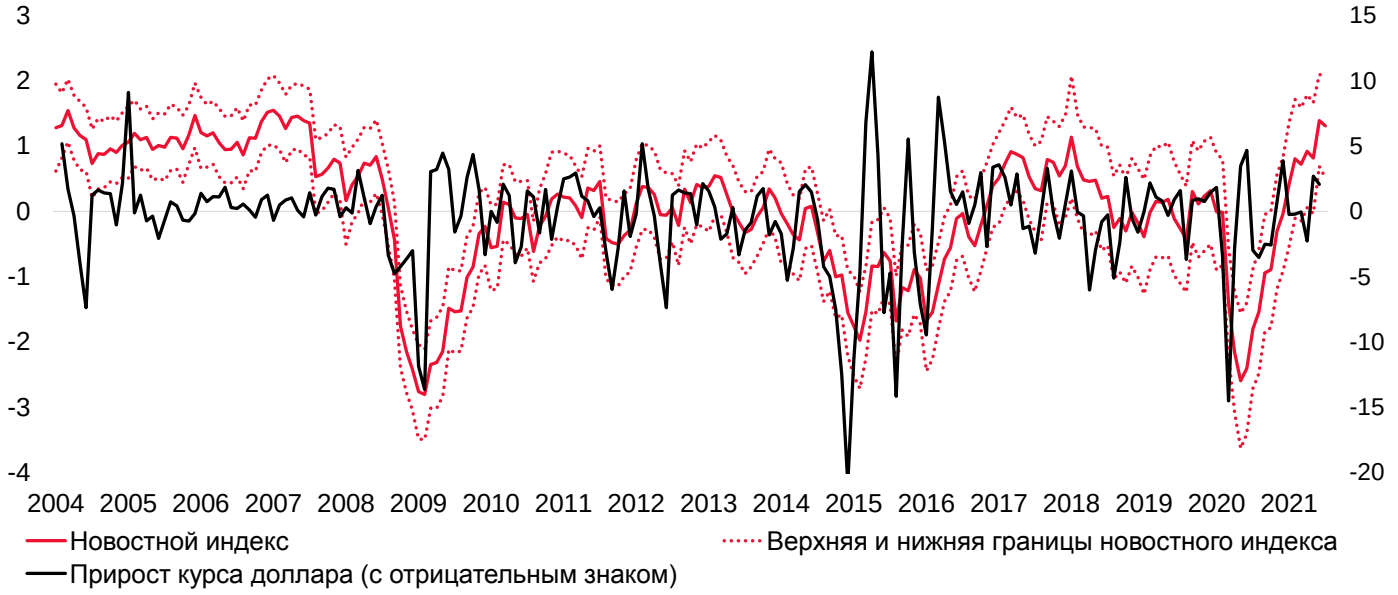


Рисунок 5. Динамика нового новостного индекса и курса рубля к доллару (м/м, %)



Приложение 1. Технические детали методологии

На первом этапе на корпусе из доступных новостей обучается LDA-модель (Latent Dirichlet Allocation, Blei et al. (2003), которая используется для выделения тем и определения вероятности каждой темы в документе. Предварительно новости лемматизируются и очищаются от слишком часто и слишком редко встречающихся слов. Для обучения используется стохастический алгоритм, похожий на алгоритм описанный в Hoffman et al. (2010) и Hoffman et al. (2013) с небольшой модификацией, которая введена ввиду неравномерного распределения количества новостей по источникам. В каждый период и для каждого месяца мы формируем выборку таким образом, чтобы количество новостей для каждого из присутствующих источников было одинаковым и равным минимальному по количеству новостей в этом месяце источнику[5].

Далее из 100 обученных тем, выделяется та, которая в наибольшей степени отражает макроэкономическую тематику (ключевые слова для этой темы изображены на Рисунке 2) и каждой новости приписывается вероятность (доля) этой темы на основе обученной ранее LDA модели[6].

На третьем этапе рассчитываются индивидуальные индексы по отдельным источникам новостей путем усреднения внутри месяца произведения вероятности макроэкономической темы на тональность новости:

$$I_t^k = \frac{1}{N_t^k} \sum_{i=1}^{N_t^k} p_t^{i,k} S_t^{i,k}$$

где I_t^k – значение индивидуального индекса *k*-го источника в месяц t;

N_t^k – количество новостей, доступных для *k*-го источника в месяц t;

$p_t^{i,k}$ – вероятность макроэкономической темы в *i*-ом документе *k*-го источника в месяц t;

$S_t^{i,k}$ – тональность новости в *i*-ом документе *k*-го источника в месяц t, рассчитываемая как:

$$S_t^{i,k} = \frac{T_{+,t}^{i,k} - T_{-,t}^{i,k}}{L_t^{i,k}}$$

где $T_{+,t}^{i,k}$ и $T_{-,t}^{i,k}$ – суммарная тональность позитивных и негативных слов в *i*-ом документе *k*-го источника в месяц t, определяемая на основе словаря;

$L_t^{i,k}$ – количество слов в *i*-ом документе *k*-го источника в месяц t.

Таким образом, увеличение доли макроэкономической тематики в *i*-ой новости ($p_t^{i,k}$) и поляризация тональности ($S_t^{i,k}$) увеличивает влияние этой новости на индивидуальный индекс.

---

[5] Мы используем реализацию LDA, реализованную в библиотеке Gensim (Rehurek and Sojka (2010)).

[6] Если быть точнее, то в качестве вероятности тем используется среднее вариационной аппроксимации апостериорного распределения тем каждого документа.

Для агрегации индивидуальных индексов в финальный новостной индекс подобно Stock and Watson (1991) используется линейная модель пространства состояний (SSM, State Space Model) с одним скрытым фактором. Модель оценивается на выборке с января 2004 по декабрь 2019 года с помощью метода максимального правдоподобия. Формально ее можно записать в следующем виде:

$$I_t^k = a^k + \lambda^k I_t + \varepsilon_t^k, \qquad k = 1,2,3$$

$$\varepsilon_t^k \sim N\left(0, (\sigma^k)^2 + \left(\sigma_t^k\right)^2\right), \qquad k = 1,2,3$$

$$I_t = bI_{t-1} + e_t$$

$$e_t \sim N(0,1)$$

где I_t – новостной индекс в момент времени t;

ε_t^k – ошибка наблюдений для уравнения наблюдений индивидуального индекса *k*-го источника в месяц t;

e_t – ошибка в AR(1) процессе уравнения состояний в момент времени t;

$a^k, \lambda^k, b, \sigma^k$ – оцениваемые коэффициенты модели;

σ_t^k – стандартное отклонение индивидуального индекса *k*-го источника в месяц t. Параметр σ_t^k не оценивается в модели и индивидуален для каждого индивидуального индекса в каждый из рассматриваемых месяцев. Он введен для того, чтобы в случае малого количества новостей конкретного источника или же сильной их несогласованности уменьшить вклад данного источника в финальный индекс. Этот параметр посчитан как стандартное отклонение оценки I_t^k (стандартное отклонение вклада одной новости скорректированное на корень из количества новостей).

Граничные значения индекса построены как 3 стандартных отклонения новостного индекса из SSM модели при всех зафиксированных параметрах (без использования обратного сглаживания). Стоит отметить, что не нужно интерпретировать данные границы как доверительные интервалы ввиду того, что они не включают в себя несколько видов неопределенности, например, неопределенность коэффициентов при построении LDA- и SSM-моделей. Эти границы (и в особенности верхняя) выбраны исходя из визуального анализа исторической динамики индекса в кризисные эпизоды и призваны быть скорее сигналом значимого негативного/позитивного новостного фона, чем доверительными интервалами в их классическом понимании.

Шкала итогового индекса была выбрана таким образом, чтобы его стандартное отклонение на периоде обучения модели (с января 2004 по декабрь 2019 года) было равным единице.

Список литературы

Blei D., Ng A. and. Jordan M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3 (2003), 993-1022.

Hoffman M., Blei D., Bach F. (2010). Online Learning for Latent Dirichlet Allocation.

Hoffman M., Blei D., Wang C., Paisley J. (2013). Stochastic Variational Inference. Journal of Machine Learning Research, 14 (2013), 1303-1347. In Neural Information Processing Systems.

Rehurek R., Sojka P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Stock J., Watson M. (1991). A Probability Model of the Coincident Economic Indicators. In: Moore. G and Lahiri K., The Leading Economic Indicators: New Approaches and Forecasting Records. Cambridge University Press. 63-90.

Авторы: Сергей Селезнев, Денис Шибитов, Сергей Власов, Дмитрий Чернядьев