Новый подход к определению отношения авторов коротких текстов к обсуждаемым темам на примере оценки инфляционных ожиданий

© М. Л. Андреев
Московский государственный университет имени М. В. Ломоносова,
Москва
mark.andreev@gmail.com

Аннотация. Предложен новый подход к измерению инфляционных ожиданий российского населения на основе активности населения на официальных сайтах финансово-ориентированных СМИ и их страницах в социальных сетях. Комментарии, представляющие собой короткие тексты, предварительно автоматически фильтруются на соответствие отношения к теме «инфляция» с помощью ключевых слов, составленных экспертом, и далее подвергаются анализу с помощью методов машинного обучения. Рассмотрены вычисляемые свойства отобранных комментариев, проведено тематическое моделирование с помощью методов вероятностного тематического моделирования для анализа главных тем, содержащихся в отобранных комментариях. Данный подход позволяет получать высокочастотную, экономически адекватную и обоснованную оценку инфляционных ожиданий населения РФ.

Ключевые слова: инфляционные ожидания, машинное обучение, анализ текстов, анализ естественного языка, тематическое моделирования.

A New Approach to Determining the Attitude of Authors of Short Texts to the Topics Discussed in the Texts on the Example of Estimating the Inflations Expectations

© Mark Andreev
Lomonosov Moscow State University,
Moscow
mark.andreev@gmail.com

Abstract. The paper suggests a new approach to measuring inflation expectations of the the Russian population based on its activity on official websites of financial-oriented mass media and their pages in social networks. Comments were previously automatically filtered to match the relationship to the topic "inflation" using keywords defined by the expert. Then, resulting set of comments was analyzed using machine learning methods. Simple calculated properties of the selected comments are considered; subject modeling is carried out using probabilistic thematic modeling methods to analyze the main topics contained in the selected comments. This approach makes it possible to obtain a high-frequency, economically adequate and justified estimate of inflation expectations of the Russian population.

Keywords: inflation expectations, machine learning, text analysis, natural language analysis, thematic modeling.

1 Введение

Под короткими текстами будем понимать комментарии пользователей в социальных сетях и на страницах официальных ресурсах СМИ. Короткие тексты характеризуется малым числом тем. Каждый комментарий имеет автора, время публикации и ссылку на статью, к которой он был оставлен. В зависимости от источника автор может иметь не только имя, но и другие атрибуты, например, место проживания, информацию о социальном графе и информацию о сообществах, в которых он состоит.

Под темой будем понимать совокупность слов, образующих смысловую повестку. При этом одна большая тема может включать в себя меньшие. Для

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/RCDL'2017), Москва, Россия, 10–13 октября 2017 года

выделения тем предлагается использовать два подхода: поиск темы по ключевым словам и с использованием тематического моделирования.

Выделение темы по ключевым словам требует формирования регулярного выражения. Такой подход ограничен списком тем, которые описал исследователь. Его результатом является список комментариев, относящихся к заданной теме. Для дальнейшего исследования полученных комментариев предлагается использовать тематическое моделирование.

Тематическое моделирование позволяет выделить из текстов заданное заранее количество тем. Уменьшая количество тем, исследователь будет получать более глобальные темы, увеличивая их количество, будет получать подтемы больших тем [4].

Под отношением к теме будем понимать как факт упоминания данной темы пользователем, так и эмоциональную окраску сообщений, в которых упоминается тема.

Оценка тональности комментариев возможна как по средствам заранее сформированных словарей, содержащих информацию о тональности каждой словоформы [7], так и с помощью методов машинного обучения, требующих размеченных текстов на предмет их тональности [6]. В данной работе рассматривается второй подход, не требующий кропотливой работы лингвиста.

Таким образом, предлагается фильтровать исходную совокупность комментариев на отношение к исследуемой теме, представленной в виде регулярного выражения, а затем исследовать статистики вычислимых свойств комментариев: количество упоминаний темы в единицу времени, количество эмоционально окрашенных комментариев в единицу времени. Для исследования тем, являющихся частью исследуемой, предлагается использовать тематическое моделирование.

Для иллюстрации работы предложенного подхода рассмотрим задачу измерения инфляционных ожиданий населения РФ на основе его активности на официальных сайтах финансово-ориентированных СМИ и их страницах в социальных сетях.

В экономике инфляционными ожиданиями называют предполагаемые уровни инфляции, основываясь на которых производители и покупатели строят свою будущую ценовую и кредитно-финансовую политику [1]. Влияние на инфляционные ожидания оказывает ЦБ РФ в рамках режима инфляционного таргетирования. Инструментом воздействия с сентября 2013 г. является ключевая ставка. Кроме того, среди косвенных инструментов воздействия на инфляционные ожидания ЦБ РФ использует информационную политику, постоянно объясняя населению свои действия и дальнейшие планы. Традиционный метод оценивания инфляционных ожиданий подразумевает проведение опросов. Используя данный подход, агентство ООО «ИНФОМ» оценивает инфляционные ожидания населения РФ. Главные недостатки такого подхода состоят в низкой частоте обновления индекса (раз в месяц); ограниченности выборки, которая состоит всего из 2000 домохозяйств; скорости публикации индекса (результаты опросов ООО «ИНФОМ» запаздывают примерно на две недели после проведения опросов за счет необходимости обработки полученных данных); отсутствие возможности пересчета показателей при изменении методологии опросов высоких издержках при построении.

Подход, предлагаемый в данной статье, лишен вышеупомянутых недостатков. Высокая частота обновления индекса обеспечивается автоматизацией сбора данных и последующим анализом на высокопроизводительном вычислительном кластере. Выборка ограничена только количеством пользователей, комментирующих новости в отобранных источниках. Сохранение потока сообщений в специальное хранилище позволяет обновить индекс при изменении методологии, без повторного сбора данных.

Чувствительность индекса к изменениям инфляционных ожиданий населения отчетливо видна на графике его изменения – по ситуации в конце 2014 – начале 2015 годов. Этот период характеризуется всплеском волнения населения. Экономическое обоснование полученного индикатора детально рассмотрено в [1].

2 Подход к построению системы

2.1 Концептуальное описание

Весь процесс построения индекса инфляционных ожиданий населения можно разделить на два логических этапа: сбора данных и анализа полученных данных. Общим для двух этапов может быть хранилище, в которое поступают данные из модуля сбора данных, а затем обрабатываются модулем анализа данных. Результат работы модуля анализа данных сохраняется в хранилище отчетов.

Основываясь на данной концепции, рассмотрим два варианта организации такой системы. В первом случае система будет сохранять собранные данные в базу данных, во втором – отправлять данные в очередь, из которой модуль анализа данных будет забирать сообщения.

Для реализации концептуального прототипа был выбран первый вариант, подразумевающий последовательную работу модулей: вначале собираются все данные, потом анализируются. Данный подход имеет более простую реализацию, чем второй, а также позволяет оперировать сразу со всей выборкой данных.

Второй подход, основанный на очереди сообщений, подразумевает батчевую обработку данных на лету. Такой подход более производителен: большая скорость обработки данных и отсутствие требования хранения всех данных в оперативной памяти.

Рисунок 1 Концептуальная схема системы обработки данных

Первый метод будем называть офлайновым, второй – онлайновым, исходя из скорости обработки поступающих данных.

2.2 Система с офлайн обработкой данных

Рассмотрим подробнее первый способ организации приложения. Модуль сбора данных состоит из «краулера», собирающего содержимое интернет-ресурсов и очереди задач, в которую «краулер» помещает ссылки на станицы, которые он планирует посетить в дальнейшем. Загруженные страницы программа сбора сохраняет в формате json в NoSQL СУБД MongoDB. Использование данной базы продиктовано необходимостью хранить структуры данных, содержащие вложенные поля и имеющие непостоянную структуру.

Особенностью данного модуля является поддержка режима распределенного сбора данных за счет наличия внешней очереди задач, реализованной с помощью сервера очередей RabbitMQ.

Для реализации «краулера» использовался язык Java, с помощью которого было построено приложение, использующее многопоточные возможности языка для ускорения сбора данных. Отказ от выбора готового решения был обусловлен необходимостью собирать данные из неоднородных источников, что требует персонального подхода к извлечению данных.

Рисунок 2 Схема реализации системы с офлайн обработкой данных

Модуль анализа данных был написан на языке Python с использованием библиотек анализа данных Pandas, Matplotlib, Scikit-learn. Исходный код исследований хранился в формате блокнотов Jupyter. Для изоляции окружения использовался Docker, такой подход позволил добиться воспроизводимости результатов, несмотря на возможные обновления рабочей системы, влияющие на реализацию алгоритмов машинного обучения.

Для вторичного анализа данных, полученный от основного модуля обработки данных, использовался язык R, исходные коды которого так же хранились в формате блокнотов Jupyter. Использование данного инструмента вызвано его популярностью в среде аналитиков, как следствие большего разнообразия модулей для статистического анализа, чем в экосистеме Python.

2.3 Система с онлайн обработкой данных

Рассмотренная ранее система наиболее оптимальна для построения исследовательского прототипа, призванного проверить базовую гипотезу. Однако для использования системы в условиях высокой нагрузки – большого потока данных от множества «краулеров» – данная система плохо пригодна. По этой причине предлагается рассмотреть потоковую обработку данных, подразумевающую отправку данных в очередь, а не напрямую в базу данных. Из очереди сообщений должны формироваться «батчи» данных, которые следует отправлять в систему распределенной обработки данных, например, Apache Spark [3]. Одновременно с этим стоит сохранять данные в специальные долгосрочные хранилища, имеющие пониженную цену на хранение данных по сравнению со стандартными облачными хранилищами. Примером долгосрочного хранилища является Amazon Glacier, Azure LRS. Архивирование данных позволяет воспроизвести вычисления, полученные ранее.

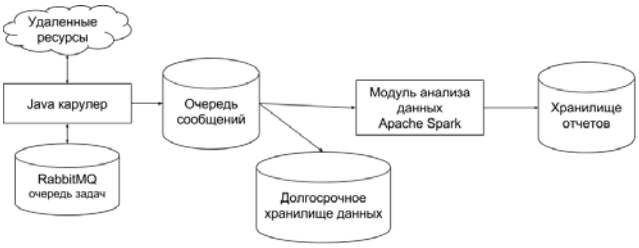

Рисунок 3 Схема реализации системы с онлайн обработкой данных

3 Построение индекса инфляционных ожиданий

Для построения индекса инфляционных ожиданий использовались не все комментарии, а лишь относящиеся к теме «инфляция». Для фильтрации целевых комментариев использовались регулярные выражения, составленные экспертом-экономистом. В анализе оценивалось как абсолютное число целевых комментариев в единицу времени, так и их вычисляемые свойства: эмоциональный окрас, тематика. Тональность комментария можно рассматривать как отношение пользователя к проблеме, рассматриваемой в статье, к которой был оставлен комментарий. Абсолютное число эмоционально окрашенных и нейтральных комментариев в единицу времени оказалось коррелированным с индексом, предоставляемым ООО «ИНФОМ».

3.1 Оценка тональности комментариев

Для оценки тональности комментариев использовались методы машинного обучения, в частности

логистическая регрессия, метод опорных векторов. Для обучения классификаторов использовалась размеченная выборка, состоящая из русскоязычных сообщений твиттера [2]. Классификатор решал задачу бинарной классификации разделения комментариев на классы «негативный» и «позитивный». На основе вероятности отношения к классу «негативный», предоставляемой обученным классификатором, принималось решение об отнесении комментариев к трем классам «позитивный», «нейтральный», «негативный». Дискретизация проводилась на основе принадлежности к полуинтервалам и отрезку: [0, 0.25) [0.25, 0.75] (0.75, 1].

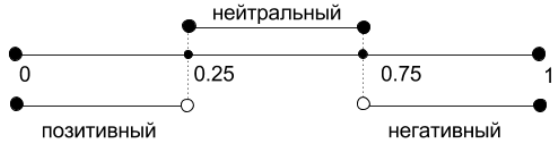

Рисунок 4 Дискретизация вероятности отнесения к классу «негативный комментарий»

Ниже представлены результаты оценки качества классификации различных моделей и методов предобработки данных, вычисленные на основе выборки сообщений из русскоязычного твиттера [2] с помощью метода перекрестной проверки с разбиением на 10 частей.

Таблица 1 Оценка качества моделей

Алгоритм	Предобработка	Верность,%
Лог регрессия	Мешок слов	76.7
	TF-IDF	76.0
Метод опорных векторов	Мешок слов	75.0
	TF-IDF	76.3

Полученная модель, вычисляющая количество эмоционально окрашенных комментариев, была противопоставлена результатам индикатора на основе опросов (традиционных подход).

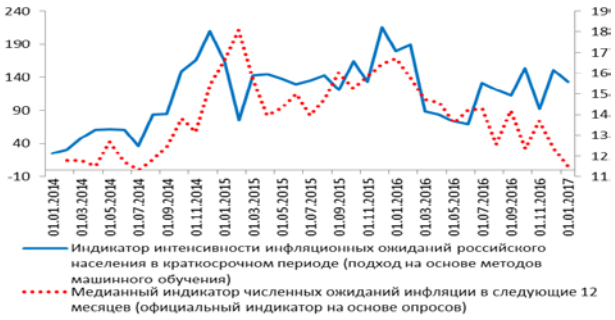

Рисунок 5 Классический индикатор и индикатор на основе методов машинного обучения

Из Рис. 4 видно, что полученный индикатор опережает индикатор, полученный традиционным путем.

3.2 Тематическое моделирование комментариев

Для анализа содержания комментариев было предложено использовать вероятностное тематическое моделирование. Такой подход позволил нам рассмотреть темы и их долю в различные периоды времени, избежав чтения всех комментариев экспертом. Вместо этого предстоит лишь рассмотреть небольшое число тем. Каждая тема представлена ключевыми словами, характеризующими ее. Для использования математических моделей тематического моделирования требуется предварительно обработать текст: нормализовать слова и удалить стоп слова и спец. символы. Для решения данной задачи использовалась библиотека NLTK и pymorphy2.

В качестве инструмента для вероятностного тематического моделирования был использован BigARTM [4], который реализует модуль ARTM (Аддитивная регуляризация тематических моделей) в качестве математической модели взаимодействия документов, терминов и тем.

Классические модели тематического моделирования малоинтерпретируемы на коротких текстах. По этой причине использовалось нестандартное представление документов – WNTM [5], которое рассматривает взаимную встречаемость слов: для каждого слова рассматривается его локальный контекст.

Количество тем для построения модели, настраивается пользователем самостоятельно. При этом увеличение числа тем ведет к слиянию менее популярных тем в одну.

На данный момент авторами не выработана окончательная методика визуализации тем с привязкой ко времени. Концептуальное видение решение данной проблемы изображено на Рис. 6. График отражает общий интерес к теме, выражающийся в абсолютном количестве сообщений, содержащих паттерн темы (удовлетворяют регулярному выражению). На оси абсцисс отложены даты, на оси ординат – абсолютное количество сообщений. Окружность, расположенная в один из моментов времени, демонстрирует момент времени, в окрестности которого производится детализация подтем посредством тематического моделирования. Список тем, характеризующийся ключевыми словами, изображен в рамке под графиком.

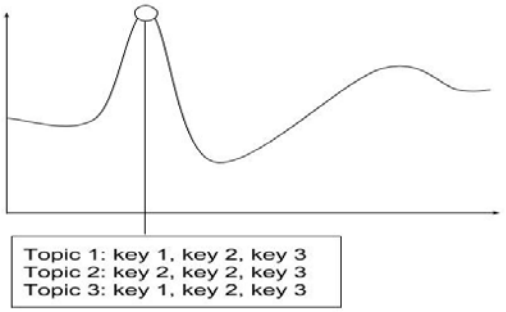

Рисунок 6 Концепция визуализации тематической модели

Авторы также планируют разработать подход для визуализации тем с учетом географии пользователей, оставляющих комментарии.

4 Заключение

Разработан подход для получения высокочастотной оценки инфляционных ожиданий населения РФ, который был реализован в виде прототипа системы анализа комментариев. Полученные практические результаты теоретически обоснованы и опубликованы в тематическом журнале [1].

Для решения данной задачи был самостоятельно реализован модуль сбора данных, имеющий специализированные компоненты извлечения данных для конкретных источников данных. Особенностью этого модуля является возможность организации распределенного сбора информации. Исследовательский код, расположенный в модуле анализа данных, использовал готовые библиотеки анализа данных. Авторы видят возможность реализации собственной тематической модели на базе BigARTM, которая бы учитывала дополнительные факторы, например, географию пользователей.

Представленный подход к построению индикатора позволяет расширять список источников комментариев, а также внести изменения в модуль анализа данных уже в процессе эксплуатации системы, уточняя значение индекса посредствам повторного вычисления с использованием новых данных.

Область применения данного подхода может выходить за рамки оценки инфляционных ожиданий и использоваться для отслеживания интереса пользователей к различным темам. Примером таких тем может быть отношение пользователей к коммерческим организациям или публичным лицам. Отслеживание интереса позволит оперативно информировать PR агентства об имидже клиента.

Подход к построению системы, представленный в статье, позволяет адаптировать систему для работы с большими данными. Увеличение охвата пользователей приведет к более точной оценке отношения пользователей к теме.

Авторами не решена проблема визуализации тематической модели для комментариев с учетом времени их публикации. Планируется модифицировать метод фильтрации комментариев, который бы учитывал место проживание пользователя. Отдельной подзадачей является идентификация ботов среди комментаторов для исключения их из рассмотрения, либо выделения в отдельную группу для информирования исследователя об их наличии.

Поддержка. Работа выполнена при поддержке РФФИ (грант 16-07-01028).

Литература

[1] Голощапова, И., Андреев, М.: Оценка инфляционных ожиданий российского населения методами машинного обучения. Вопросы экономики, (6), сс. 71-93. Некоммерческое партнерство «Редакция журнала «Вопросы экономики»» (2017)

[2] Рубцова, Ю. Построение корпуса текстов для настройки тонового классификатора. Программные продукты и системы, (1), сс. 72-78. Научно-исследовательский институт «Центрпрограммсистем» (2015)

[3] Клеменков, П. Пайплайн машинного обучения на Apache Spark. Конференция HighLoad++ (2016)

[4] Vorontsov, K. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. Int. Conf. on Analysis of Images, Social Networks and Texts, pp. 370-381. Springer International Publishing (2015)

[5] Zuo, Y., Jichang Z., Ke X. Word Network Topic Model: a Simple But General Solution for Short and Imbalanced Texts. arXiv preprint arXiv:1412.5404 (2014)

[6] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment Classification using Machine Learning Techniques. Proc. of the ACL-02 Conf. on Empirical Methods in Natural Language Processing, 10, pp. 79-86. Association for Computational Linguistics (2002).

[7] Rao,. Y.: Building Emotional Dictionary for Sentiment Analysis of Online News. World Wide, (4), pp. 723 (2014)

[8] Chang, J. Reading tea leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems, pp. 288-296 (2009)