Компаративный анализ методологических основ задач прогнозирования исходов выборов и протестов по цифровым следам пользователей в социальных медиа♣

А.А. Фильченков, А.А. Азаров, М.В. Абрамов

МГГУ им. Шолохова, Университет ИТМО, СПИИРАН
aaafil@mail.ru, artur-azarov@yandex.ru, mva16@list.ru

Аннотация

Задача прогнозирования протестной активности по цифровым следам пользователей социальных медиа до сих пор является нерешенной, кроме того в широком доступе не существует научных статей на эту тематику. В данной статье использован сравнительный анализ этой задачи и задачи прогнозирования результатов выборов, являющейся уже достаточно изученной. Сравнение осуществлялось по шести выделенным критериям: представленность, репрезентативность, открытость, теоретическая предсказуемость, математическая постановка задачи и доступность данных.

1. Введение

Задача предсказания (прогнозирования) является фундаментальной научной задачей. Осуществление предсказаний возникновения и течения событий и явлений общественной и политической жизни является одной из фундаментальных задач политологии.

Увеличение роли информационных технологий в общественной жизни сопровождается в первую очередь массовым вовлечением населения в онлайн-коммуникацию, в первую очередь в социальных медиа. Это приводит к появлению обилия цифровых следов (сообщений, картинок, плафтормозависимых взаимодействий, таких как «лайки», «репосты»). Рядом исследователей была показана принципиальная возможность прогнозирования событий на основе обработки цифровых следов в социальных медиа [1].

Задача прогнозирования результатов выборов по социальным медиа считается достаточно хорошо изученной и данной тематике посвящены десятки работ. Задача прогнозирования протестов по социальным медиа в научной литературе, напротив, не освещается. В данной работе будет проведен компаративный анализ методологических основ двух указанных задачи, на основе чего будут указаны потенциальные трудности и ограничения, связанные с прогнозированием протестной активности.

2. Прогнозирование результатов выборов

Выделим наиболее часто используемые подходы к прогнозированию результатов выборов по социальным медиа.

Одной из популярных методик является использование профиля настроения (Profile of Mood States), который строится для каждого пользователя социальной сети [14, 8, 31, 13, 26]. В рамках этого направления реализовано несколько различных моделей, различающихся по сложности и учитываем признакам. Наиболее продвинутой моделью является набор из скрытых составных моделей, индивидуальной для каждого избирателя, в репрезентативной выборке населения избирательного возраста.

Другим направлением является прогнозирование на основе различных лингвистических методик, основанных на анализе тональности высказываний [6].

Кроме того, существует множество различных концептуальных методов прогнозирования результатов выборов, таких как использования методов предсказания погоды [28]; основанных на графах социальных связей пользователей (Facebook, Twitter, YouTube, Google); на их политических предпочтениях [19, 12].

Более детальный обзор представлен в [21]

3. Два типа протестных событий

Прежде чем приступать к сравнительному анализу, необходимо ответить, что различные типы протестов принципиально различаются, что существенно влияет на возможность осуществления предсказания таких событий.

В целях настоящей работы мы выделим два типа протестов, которые условно назовем **стихийные** и **организованные**.

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

Стихийные протесты (например, события в Бирюлово в 2012 году) отличаются тем, что протест возникает спонтанно как ответ на какой-либо информационный повод (обычно, происшествие), очаг возникновения локализуется вне социальных сетей, которые, хотя и могут выступать инструментом для координации и мобилизации, не играют существенной роли в моменте образования протеста. Стихийные протесты в данном определении близки к бунтам.

Организованные протесты (например, выступления на Болотной площади), напротив, организуются преимущественно в социальных сетях, о них известно заранее, перед их проведением в сети присутствует достаточно большое число цифровых следов. Название не подразумевает, что у подобных протестов должен быть непосредственный организатор, однако их развертывание характеризуется плавностью и нарастанием, в отличие от взрывного характера стихийных протестов.

4. Сравнительный анализ

Поскольку прогнозирование выборов по данным, извлеченным из социальных медиа, уже является активно развивающейся темой, в которой получены серьезные успехи, в данной секции будут представлены результаты компаративного анализа этой задачи и задачи прогнозирования протестов. В данной статье рассмотрены шесть критериев сравнения предсказания результатов выборов и прогнозирования протестной активности.

4.1. Представленность

Критерий: для осуществления предсказания по социальным медиа необходимо, чтобы соответствующая информация была в них представлена.

Выборы: Социальные медиа используются для самопрезентации [5]. Люди выражают свое мнение на различные политические вопросы, политические концепции, о различных политических организациях и политических деятелях. Таким образом, информация представлена.

Организованный протест: Использования социальных медиа во время различных протестных акций в 2013 году было активно изучено исследователями, например, российские протесты после выборов в Думу в 2011–2012 [32, 35] и Украинский Евромайдан [2, 3, 7]. Как и в случае выборов, представлены политические взгляды, предпочтения высказывания. Более того, в социальные сети используются для мобилизации участия в протестах.

Стихийный протест: В сети представлены лишь характеристики, которые выступают латентными для определения участия в стихийном протесте.

Результаты: лучше всего информация представления на организованного протеста, хуже всего — для стихийного протеста.

4.2. Первая страница

Критерий: точности прогноза способствует репрезентативность пользователей социальных медиа [20].

Выборы: При прогнозировании результатов выборов обнаруживается серьезная проблема с репрезентативностью социальных медиа. Наблюдается серьезное смещение представленных мнений из-за «цифрового разрыва» (digital divide). Наиболее явно это проявляется по возрастному признаку: чем старше человек, тем меньше вероятность того, что он является активным пользователем [10]. Нет никаких свидетельств в пользу того, что возраст избирателей имеет такое же смещение.

С другой стороны, активность в социальных сетях и политическая активность коррелируют [34].

Организованные протесты: в основном в протестных акциях в основном участвует образованная городская молодежь [17], которая одновременно является наиболее активными участниками социальных сетей.

Более того, в [33] показано, что положительная корреляция существует также и между политической активностью в социальных сетях и политической активностью в жизни, к которой скорее относится политический активизм, чем участие в голосовании.

Стихийные протесты: в стихийных протестах участвует более возрастная аудитория, чем в организованных протестах. Предполагается также, что и менее образованная. Однако найти каких-либо исследований, освещающих состав участников подобных протестов авторам не удалось.

Результаты: больше всего свидетельств с пользу репрезентативности выделено для организованных протестов, для стихийных протестов нет релевантной информации, а для выборов заведомо известно, что репрезентативность отсутствует.

4.3. Открытость

Критерий: для возможности осуществлять предсказания требуется открытость публичного обсуждения вопросов соответствующей области [25].

Поскольку политические вопросы являются основной целью цензуры, все случаи сильно зависят от степени свободы слова в государстве и обществе. В последние годы в России наряду с другими странами, наблюдается усиление контроля над онлайн дискурсом путем ужесточения законодательной и правоприменительной практики, что, естественно, затрудняет публичную политическую дискуссию, в особенности связанную с протестной деятельностью.

С другой стороны, свобода слова традиционно рассматривается как элемент демократического общества, и через этот предиктор положительно коррелирует с коэффициентом детерминации

исхода выборов от общественных настроений. Проще говоря, степень свободы слова тесно связана с уровнем административного вмешательства в итоги голосования, что существенно затрудняет осуществление прогнозов (по социальным медиа).

Результаты: для всех случаев открытость влияет сопоставимо; ущемление свободы слова опосредованно влияет на возможность прогнозирования результатов выборов.

4.4. Теоретическая предсказуемость

Если предыдущие критерии касались представления мнений в социальных сетях, то этот критерий относится к разработанности теоретических моделей. Следует отметить, что предсказания могут быть сделаны даже при отсутствии полного понимания, как функционирует система, поведение которой будет предсказано.

Критерий: наличие теоретических моделей прогнозирования.

Выборы: для предсказания выборов предложены различные теоретические модели [11, 35].

Протесты: вопрос о том, является ли протест случайным, является дискуссионным, и в научной литературе можно найти как сторонников этой точки зрения [16, 27], так и противников, утверждающих, что он хотя бы в некоторой степени закономерен [23, 29, 22].

Наиболее известная теоретическая модель прогнозирования протестов предложена в [24], развивалась в [20, 29]. Другие популярные модели, рассматривающие протест как проявление общественного конфликта, описаны в [18, 4].

Указанные модели применимы для предсказания обоих выделенных типов протестов.

Подробный обзор теоретических моделей приведен в [9].

Результаты: вопрос о предсказуемости всех типов рассматриваемых событий является открытым, однако во всех случаях предложены теоретические модели.

2.4. Математическая постановка задачи

Критерий: математическая постановка задачи предсказания, которая определяет методы, используемые для решения.

Выборы: математическая задача предсказания исхода выборов — это задача восстановления регрессии, классическая задача статистики.

Протесты: предсказание протестов разбивается на несколько задач:

- задачу предсказания длительности протеста — это задача восстановления регрессии.
- задачу предсказания численности протестующих — тоже задача восстановления регрессии.
- задача предсказания начала протеста (наиболее важная) в общем случае — это задача предсказания. В зависимости от конкретной постановки и дополнительных предположений (то есть используемой модели) возможно использование достаточно большое число подходов.

Результаты: задача предсказания протестов в математическом смысле сложнее задачи предсказания исхода выборов.

4.6. Доступность данных

Для формирования (обучения) прогностической модели, то есть применения методов машинного обучения, необходимы обучающие наборы данных. Предлагая новую модель, ее необходимо тестировать и корректировать на данных.

Поскольку информация, представленная в социальных медиа, огромна (терабайты), невозможно хранить ее всю.

Существует три варианта решения такой проблемы.

Первый состоит в том, чтобы каждый раз заново искать данные (сообщения, связи) для событий, имевших место в прошлом. Однако такой подход сопряжен со многими ограничениями, потому что множество информации с течением времени будет утрачено. В частности, задача восстановления графа социальных связей на определенный момент времени чрезвычайно трудно реализуема и представляет отдельную исследовательскую проблему.

Второй подход состоит в том, чтобы сохранять лишь необходимые данные. Однако понятие «необходимые» определяется моделью. Поэтому, решая задачу построения модели, нельзя предугадать, какие данные в итоге потребуются.

Наконец, третий подход состоит в том, чтобы обучать и тестировать модель на новых данных. Но он предполагает, что данные должны быть регулярно доступны. Это и является последним критерием сравнения.

Выборы: проводятся регулярно, они относятся к так называемым «стабильных» событий [15]. Поэтому мы можем обучать нашу модель на новых данных, зная, с какого момента их следует собирать (например, за месяц до выборов). В отношении выборов достаточно хорошо определен контекст (известны имена кандидатов, предвыборные программы и т.д.), что упрощает задачу выбора признаков для модели.

Организованные протесты: протесты относятся к «нестабильным» сообщениям — зачастую мы не знаем, когда они наступят, чтобы заранее подготовиться и настроить под них систему. Кроме того, в отношении протестов в общем случае мы не обладаем никакими эвристиками, которые бы могли облегчить поиск.

Стихийные протесты: совершенно неопределен контекст, поскольку обычно они являются ответом на какие-либо события. Достаточно сложно

провести выделение групп по похожести событий, что еще более затрудняет поиск данных.

Результаты: выборы намного превосходят протесты по доступности данных. Именно этот критерий, на наш взгляд, является определяющим для решения задачи предсказания протеста и ее сложности по сравнению с задачей предсказания выборов

6. Заключение

В работе было выделено два типа протестов: стихийные и организованные. Подводя итоги компаративного анализа, можно сделать вывод, что предсказание организованных протестов по данным социальных медиа имеет большой потенциал. Социальные сети активно используются для мобилизации протестного населения, одновременно оставляя цифровой след мобилизации, который может быть исследован. С точки зрения машинного обучения динамику мобилизации можно описать временным рядом предикторов, на основе которого строить предсказательные модели протеста: его начала, его продолжительности и его масштаба (числа вовлеченных людей). Одновременно с этим построение прогностических моделей протеста является существенно более сложной задачей в силу нестабильности самих протестов и, как следствие, нерегулярности данных, на которых эту модель можно обучить и верифицировать.

Стихийные протесты, имеют существенно меньший потенциал для осуществления предсказания. Вероятно, предсказательная способность будет ограничена лишь оценкой уровня «взрывоопасности» ситуации.

Следует также указать на ряд теоретических ограничений.

Существенной угрозой репрезентативности данных, а, значит, и предсказательной способности моделей, являются боты. В [6] было показано, что социальные сети уязвимы для ботов. Практика формирования искусственных мнений и оценок в маркетинговых или экономических целях стала общепринятой. Таким образом, это может привести к тому, что исследователь социальных медиа будет заниматься исследованием не общества, а бот-сообщества.

Существенным глобальным риском для исследований социальных медиа, в особенности связанных с политическими вопросами, является ограничение свободы слова в Интернете.

Литература

[1] Asur S., Huberman B. APredicting the future with social media. Web Intelligence and Intelligent Agent Technology (WI-IAT) // IEEE/WIC/ACM International Conference on. IEEE. . 2010. Vol. 1. P. 492–499.

[2] Азаров А.А., Бродовская Е.В., Дмитриева О.В., Домбровская А.Ю., Фильченков А.А. Стратегии формирования установок протестного поведения в сети Интернет: опыт применения киберметрического анализа (на примере Евромайдана, ноябрь 2013 г.). Часть I // Мониторинг общественного мнения. 2014. Вып. 2 (120). С. 63–78.

[3] Азаров А.А., Бродовская Е.В., Дмитриева О.В., Домбровская А.Ю., Фильченков А.А. Стратегии формирования установок протестного поведения в сети Интернет: опыт применения киберметрического анализа (на примере Евромайдана). Часть II // Мониторинг общественного мнения. 2014. Вып. 3 (121). С. 56–74.

[4] Bagozzi B. Forecasting civil conflict with zero-inflated count models. Manuscript. Pennsylvania State University, 2011..

[5] Бараш Р.Э. Интернет как средство самоактуализации и революционной самоорганизации // Мониторинг общественного мнения. 2012. №. 3. С. 100–109.

[6] Bermingham A., SmeatonA. F. On Using Twitter to Monitor Political Sentiment and Predict Election Results // Workshop on Sentiment Analysis where AI meets Psychology. (Chiang Mai, Thailand, November 13, 2011)

[7] Bohdanova T. Unexpected revolution: the role of social media in Ukraine's Euromaidan uprising // European View. 2014. P. 1–10.

[8] Bollen J., Mao H., Pepe A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena // ICWSM. 2011. P. 450–452.

[9] Brandt P.T., Freeman, J.R., Schrodt P.A. Racing horses: constructing and evaluating forecasts in political science // 28th summer meeting of the society for political methodology. 2011.

[10] Бродовская Е.В., Шумилова О.Е. Российские пользователи и непользователи: соотношение и основные особенности // Мониторинг общественного мнения. 2013. №. 3(115). С. 3–18.

[11] Campbell J.E. The science of forecasting presidential elections // Before the Vote: Forecasting American National Elections. 2000. 169–187.

[12] Castillo C., Mendoza M., Poblete B. Information credibility on twitter // Proceedings of the 20th international conference on World wide web. 2011. P. 675–684.

[13] Contractor D., Faruquie T. A. Understanding election candidate approval ratings using social media data // Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee. 2013. P. 189–190.

[14] Curry D., Cochran J., Radhakrishnan R., Pinnell J. Hierarchical Bayesian Prediction Methods in Election Politics: Introduction and Major Test // Journal of Political Marketing. 2013. Vol. 12 (4). P. 275–305.

[15] Dunn J. Modern revolutions: an introduction to the analysis of a political phenomenon. Cambridge University Press, 1989.
[16] Eckstein H. More about applied political science // PS: Political Science & Politics. 1990. Vol. 23 (1). P. 54–56.
[17] Enikolopov R., Makarin A., Petrova M., Polishchuk L. Social Media and Protest Participation: Cross-City Evidence from Russia // XV April International Academic Conference on Economic and Social Development. 2013.
[18] Fearon J.D., Laitin D.D. Ethnicity, insurgency, and civil war // American political science review. 2003. Vol. 97 (1). P. 75–90.
[19] Franch F. Election Prediction with Social Media // Journal of Information Technology & Politics. 2013. Vol. 10 (1). P. 57–71.
[20] Francisco R. A. Theories of Protest and the Revolutions of 1989 // American Journal of Political Science. 1993. P. 663–680.
[21] Gayo-Avello D. A. Meta-analysis of state-of-the-art electoral prediction from Twitter data. 2013. arXiv preprint arXiv:1206.5851.
[22] Geller D.S. The impact of political system structure on probability patterns of internal disorder // American Journal of Political Science. 1987. P. 217–235.
[23] Gurr T. R. Political protest and rebellion in the 1960s: The United States in world perspective // Violence in America: Historical and Comparative Perspectives, Rev. Ed.,(Beverley Hills: Sage, 1979).
[24] Gurr T. R., Lichbach M. I. Forecasting domestic political conflict. J. Singer and M. Wallace, To Augur Well: Early Warning Indicators in World Politics. 1979. P. 153–194.
[25] Huff D. How to lie with statistics. WW Norton & Company? 1954.
[26] Krishnamoorthy M., Miller W., Krishnamoorthy R. Evolution of choices over time: The US Presidential election 2012 and the NY City Mayoral Election. 2013. arXiv preprint arXiv:1310.1118.
[27] Kuran T. Now out of never: The element of surprise in the East European revolution of 1989 // World politics. 1991. Vol. 44 (1). P. 7–48.
[28] Lewis-Beck M. S., Stegmaier M. To improve their predictions, election forecasters should look to other disciplines like meteorology // LSE American Politics and Policy. 2014.
[29] Lichbach M. Protest: Random or Contagious? The Postwar United Kingdom // Armed Forces & Society. 1985. Vol. 11 (4). P. 581–608.
[30] Nam T. Dynamics Between Government Signal and Dissident Groups // Annual Meeting of the Midwest Political Science Association. (Chicago, April 11, 2003).
[31] O'Connor B., Balasubramanyan R., Routledge B., Smith N. From tweets to polls: Linking text sentiment to public opinion time series // ICWSM. 2010. Vol. 11. P. 122–129.
[32] Шерстобитов А.С., Брянов К.А. Технологии политической мобилизации в социальной сети "Вконтакте": сетевой анализ протестного и провластного сегментов // Исторические, философские, политические и юридические науки, культурология и искусствоведение. Вопросы теории и практики. Тамбов: Грамота. 2013. Вып. 10-1. С. 196–202.
[33] Vesnic-Alujevic L. Political participation and web 2.0 in Europe: A case study of Facebook // Public Relations Review. 2012. Vol. 38 (3). P. 466–470.
[34] Vitak J. It's complicated: Facebook users' political participation in the 2008 election // CyberPsychology, behavior, and social networking. 2011. Vol. 14 (3). P. 107–114.
[35] White S., McAllister I. Did Russia (Nearly) have a Facebook Revolution in 2011? // Social Media's Challenge to Authoritarianism. Politics. 2014. Vol. 34 (1). P. 72–84.
[36] Yu S., Kak S. A survey of prediction using social media. 2012. arXiv preprint arXiv:1203.1647.

Comparative analysis of the predict outcomes of elections and protests tasks methodological foundations on the basis of digital traces of users in social media

Filchenkov A.A., Azarov A.A., Abramov M.V.

The problem of forecasting protest activity on digital traces of social media users is still unresolved, in addition there is no scientific articles on this topic. This article uses comparative analysis of this problem and the problem of forecasting election results, which is already sufficiently studied. The comparison was carried out on six selected criteria: representation, representativeness, openness, theoretical predictability, the mathematical formulation of the problem and the availability of data.

♣ Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 14-07-00694А.