# 3. Прогнозирование событий на основе данных, полученных из Twitter

Анализ социальных сетей становится очень важным инструментом для мониторинга сообществ, предпочтений пользователей и составления прогнозов. Среди решений для социальных сетей Twitter является одним из самых распространенных микроблогов, позволяющих пользователям иметь личную ленту новостей и прикрепленных к ней подписчиков. Подписчики получают некоторые уведомления, связанные с действиями, выполняемыми пользователями, на которых они подписаны. Типичными действиями пользователей могут быть: публикация твита, комментирование, выражение «нравится»/«избранное», ретвитит.

Таким образом, твиты и ретвиты экспонируются другим пользователям Twitter, что повышает вероятность спровоцировать их интересы и реакции: ретвиты, комментарии, отметки «нравится». Некоторые из этих механизмов могут спровоцировать вирусные процессы, которые могут привести к массовому распространению твитов в сообществе пользователей. Пользователи Twitter официально идентифицируются по «@». Любой пользователь может привлечь внимание других пользователей, включив «@username» в твит. В тексте твита каждый пользователь может обратить внимание на определенные ключевые слова, называемые хэштегами, которые отмечены «#» в качестве первого символа. Например, хэштег: «#lacasadepapel» можно использовать, чтобы отметить, что твит посвящен сериалу "Бумажный дом". Цитаты и хэштеги хорошо индексируются в Twitter и могут быть найдены как основные средства вовлечения и, таким образом, используются Twitter для распространения информации среди цитируемых пользователей и сообществ, заинтересованных в подписчиках или хэштегах, соответственно.

Благодаря вышеописанным механизмам социального взаимодействия множество пользователей присоединяются и используют Twitter каждый день: не только отдельные пользователи, но и информационные агентства, государственные учреждения, продюсеры, VIP-персоны, команды, школы, муниципалитеты, правительства с целью обмена, продвижения и коммуникации. На таких основаниях Twitter используется в качестве источника информации для распространения новостей, событий и инноваций, и, таким образом, его можно использовать как инструмент для прогнозирования различных видов событий и происшествий.

Прогнозировать можно множество событий: кассовые сборы фильмов, фондовый рынок. Тем не менее, есть область, где ожидаются отдельные результаты в конкретный момент, то есть прогнозы: политика.

Очевидно, что избиратели могут быть более или менее близки к кандидату, но, в конечном счете, каждый из них должен отдать свой голос. Аналогичным образом, опросы меняются с течением времени, но как только избирательные участки закрываются, количество бюллетеней фиксируется, и определяется победитель.

Как подразумевалось выше, существуют противоположные точки зрения относительно предсказательной способности данных Twitter о выборах. Однако здесь нет противоречий: анализируя предыдущие отчеты, можно прийти к обоснованным выводам о предполагаемой силе современных методов.

## 3.1. Характеристика методов прогнозирования выборов на основе данных из Twitter

Хотя это и не указано, предполагается, что любой метод прогнозирования результатов выборов на основе данных Twitter является алгоритмом. В противном случае он был бы непрактичным и бессмысленным.

Следовательно, такие методы обрабатывают некоторую коллекцию твитов для составления прогнозов; они параметризуются для адаптации к различным сценариям и, наконец, прогнозы могут быть более или менее подробными. В итоге они должны быть оценены в сравнении с фактическими результатами.

Таким образом, существует ряд характеристик и подхарактеристик, определяющих любой метод прогнозирования результатов выборов с помощью Twitter:

1. Период и методы сбора;
2. Методы выбора данных:
   1. «Чистота» данных: просмотр твитов только от потенциальных избирателей;
   2. Послабление демографических ограничений;
   3. Дискредитация: гарантия, что любая демографическая предвзятость в пользовательской базе Twitter является ремоделирующей;
3. Методы прогнозирования:
   1. Метод определения намерений голосования по твитам;
   2. Характер вывода;
   3. Характер прогноза;
   4. Детализация;
4. Оценка эффективности.

### 3.1.1. Общая характеристика исследований, проведенных на сегодняшний день

Первое, что показывает таблица 2 - то, что литература, касающаяся прогнозирования выборов, вовсе не является предсказанием. Все отчеты были написаны постфактум, и те исследования, в которых обсуждаются многообещающие или даже положительные результаты, фактически описывают, как можно было предсказать выборы.

Хотя и происходит репликация результатов, обычно она не проводится теми же авторами, предлагающими оригинальный метод. Как будет показано, в Твиттере широко использовались только два «вариант» вывода о голосовании, а именно подсчет твитов и сентимент-анализ на основе лексики, и в ряде работ оценивались и сравнивались оба варианта.

В дополнение к этому ряд документов посвящен тем же выборам. [77] и [101] освещали президентские выборы в США в 2008 году. И Тумасьян [78], и Юнгхерр [81] освещали федеральные выборы в Германии в 2009 году. Эти документы должны прояснить, можно ли получить согласованные результаты из данных Twitter или, наоборот, результаты сильно зависят от решений, принимаемых исследователями.

Остальные статьи соответствуют сценариям отдельных случаев, но исследование Метаксаса [102] заслуживает особого внимания, поскольку охватывает шесть различных гонок на одних и тех же выборах в Соединенных Штатах. Его статья проливает свет на то, могут ли положительные результаты, достигнутые с помощью этих методов, быть результатом чистой случайности.

Что касается выборов, освещаемых в литературе, то наиболее изученным сценарием являются Соединенные Штаты (4 статьи), за ними следует Германия (2 статьи); Ирландия, Сингапур и Нидерланды освещаются по одной статье каждая.

### 3.1.2. Период сбора

Таблица 2 показывает существенные различия между исследованиями в отношении некоторых характеристик схемы.

Например, период сбора данных сильно варьируется. Таблица 1 показывает, что в некоторых исследованиях данные собирались всего за неделю до выборов, в то время как в других - за недели, месяцы или даже годы.

Однако существует консенсус относительно конечной точки сбора данных: за день до выборов. Единственной статьей, нарушающей это правило, является исследование Тумасьяна [78], и, как будет показано, это было предметом критики исследования Юнгхерра [81].

Следует отметить, что неясно, какое влияние период сбора оказывает на прогнозы. Юнгхерр [81] показал, что при использовании разных временных окон производительность претерпевала существенные изменения. Метаксас [102], используя данные всего за одну неделю, смогл получить как правильные, так и неправильные прогнозы.

Следовательно, ясно, что в этом отношении необходимы дальнейшие исследования и убедительные аргументы для выбора определенного периода сбора.

|  |  |  |
| --- | --- | --- |
| Автор | Начало сбора (до выборов) | Конец сбора |
| Livne A. [103] | За 3 года | День выборов |
| O`Connor B. [77] | За 10 месяцев | День выборов |
| Gayo-Avello D. [101] | За 5 месяцев | День выборов |
| Tumasjan A. [78] | За 7 недель | За 1 неделю до дня выборов |
| Jungherr A. [81] | За 7 недель | День выборов |
| Scoric M. [104] | За один месяц | День выборов |
| Bermingham A. [105] | За 3 недели | День выборов |
| Metaxas P. [102] | За 1 неделю | День выборов |
| Tjong K.S. [106] | За 1 неделю | День выборов |

Таблица 1 - различные периоды сбора данных, используемые в литературе, упорядоченные по убыванию продолжительности

### 3.1.3. Выбор данных

Выбор данных относится к мерам, принятым для составления прогноза путем опоры на тех пользователей, которые являются потенциальными избирателями (1), учета только тех твитов, которые касаются избирательного процесса (2), и корректировки собранных данных с учетом любых демографических отклонений в пользовательской базе Twitter (3).

Первый из них, обозначенный выше как выбор, включает в себя те решения, которые были приняты для отбора пользователей Twitter, которые, вероятно, будут голосовать на интересующих выборах. Излишне говорить, что такая информация недоступна, и, следовательно, максимум, что можно сделать, - ограничить сбор данных теми пользователями, которые находятся в интересующей области.

Это исключило бы тех пользователей, которые не указали действительное местоположение, или даже эмигрантов, имеющих право голоса, но это компромиссное решение, позволяющее избежать учета пользователей, выражающих свои взгляды на кампанию, не имея права голоса.

Такое решение возможно, используя только твиты с геолокацией или проверяя строку местоположения пользователей в коллекции. Довольно удивительно, что только в двух исследованиях в литературе применялась такая мера. Гайо-Авелло [102], полагаясь на твиты, геолокированные в интересующих округах, и Скорик [104], ограничивая набор данных теми пользователями, которые находятся в Сингапуре.

Возможно, следует согласиться с тем, что Тумасьян [78], Юнгхер [81] и Джонг [106] гарантируют в определенной степени чистоту коллекции на основе использования языка. Первые две статьи посвящены выборам в Германии, и, таким образом, предполагается, что твиты о немецких партиях и политиках, написанные на немецком языке, скорее всего, созданы немецкими пользователями, а не австрийскими или швейцарскими пользователями. Аналогичным образом, третье исследование касается выборов в Нидерландах, и данные, вероятно, получены в Нидерландах, а не в Бельгии.

Тем не менее, при рассмотрении более глобализированных языков (например, английского) такое предположение неприемлемо, и следует учитывать геолокацию, чтобы гарантировать чистоту набора данных.

Вторая мера по очистке данных - дискредитация. Это включает в себя любую последующую обработку набора данных для удаления твитов или пользователей, не имеющих отношения к избирательному процессу или не соответствующих потенциальным избирателям, соответственно. Другими словами, это подразумевает удаление спама, слухов, пропаганды, дезинформации и пользователей, создающих шумные твиты.

Таблица 2 показывает, что ни в одной статье в литературе не были приняты такие меры, хотя некоторые из них признают наличие проблемы. Например, Метаксас [102] сделал это предупреждение: «Спамеры и пропагандисты пишут программы, которые создают множество поддельных аккаунтов и используют их для интенсивного твиттера, усиливая свое сообщение и загрязняя данные для любого наблюдателя. Известно, что такое случалось в прошлом. Вполне разумно, что, если изображение, представленное социальными сетями, важно для кого-то (рекламодателей, спамеров, пропагандистов), скорее всего, найдутся люди, которые попытаются его подделать».

Более того, они провели эксперимент, чтобы проверить устойчивость обычно применяемых методов анализа настроений к такого рода манипуляциям, обнаружив, что: «Просто полагаясь на полярную лексику, тонкости пропаганды и дезинформации не только упускаются из виду, но даже неправильно интерпретируются».

Следовательно, меры по фильтрации данных в политических твитах не являются необязательными для получения точных прогнозов.

Третья и последняя мера по очистке данных - послабление. Пользовательская база Twitter не является репрезентативной выборкой населения, и эту проблему можно решить, определив, к каким демографическим слоям принадлежат пользователи, и соответствующим образом взвесив их твиты.

Низкая репрезентативность Twitter широко обсуждалась. Например, Бойд [107] написала следующее: «Большие данные открывают новые возможности для понимания социальной практики. Конечно, следующее утверждение должно начинаться с «но». И это «но» просто: [...] только потому, что у вас большое N, не означает, что оно репрезентативно».

Кроме того, даже в Соединенных Штатах использование Twitter незначительно (11% американцев), а их пользователи «в подавляющем большинстве молоды» [108].

Эта низкая репрезентативность является серьезной проблемой, поскольку доминирующие демографические группы могут склоняться к нескольким избранным политическим вариантам [109], и такой уклон сильно искажает результаты [101].

Следовательно, единственный способ бороться с этой проблемой - определить как можно больше демографической информации о пользователях Twitter (1) и взвесить твиты из каждой группы в соответствии с предварительной информацией об их участии в выборах (2).

Излишне говорить, что первая задача далеко не из легких. В отличие от других сервисов, таких как Facebook, профили Twitter не содержат структурированной информации. Нет способа указать пол или возраст пользователя, и вместо этого профили состоят из свободных текстовых полей для имени, местоположения, веб-сайта и биографии.

Тем не менее, это не является неразрешимым, и в более позднем разделе комментируются некоторые ссылки по этому вопросу. Действительно, таблица 2 показывает, что в двух статьях была предпринята попытка дискредитации.

Гайо-Авелло [101] смог получить возраст примерно 2500 пользователей в своем наборе данных, сопоставив их полные имена и округ с общедоступными онлайн-записями. Таким образом, он обнаружил, что в наборе данных преобладали пользователи в возрасте от 18 до 44 лет. Затем, взвесив их твиты в соответствии с возрастом участников выборов 2004 года, он смог уменьшить ошибку с 13,10% до 11,61% – значительный прирост производительности.

Джонг [106] попробовал другой подход: искажение данных в соответствии с предполагаемыми политическими пристрастиями населения. Безусловно, такая функция чрезвычайно актуальна, особенно если подозревается «застенчивый» эффект. К сожалению, их результаты были неубедительными, поскольку авторам приходилось полагаться на данные предвыборных опросов, которые можно было рассматривать как переобучение (1), и эффективность метода при обесценивании была не лучше, чем у более простого метода, основанного на количестве твитов (2).

Таким образом, искажение данных Twitter в соответствии с демографическими характеристиками пользователей представляется не только неизбежным, но и положительно влияет на производительность.

### 3.1.4. Методы прогнозирования

#### 3.1.4.1. Подсчет голосов путем подсчета твитов

Для определения голосов по твитам использовались два основных метода. Первый, первоначально предложенный Тумасьяном [78], состоит из простого подсчета твитов с упоминанием данного кандидата или партии. Чем больше количество твитов, тем выше процент голосов.

Такой метод привлекателен по многим причинам: он прост в реализации, его можно применять практически в режиме реального времени, и его можно использовать как для получения агрегированных показателей голосов, так и для определения намерений отдельных лиц голосовать. Более того, Тумасьян утверждал, что метод продемонстрировал хорошую производительность: «Простое количество твитов отражает предпочтения избирателей и приближается к традиционным предвыборным опросам».

Таблица 2 показывает, что они сообщили об ошибке в 1,65% на федеральных выборах в Германии в 2009 году.

Юнгхер [81] позже раскритиковали некоторые решения, принятые Тумасьяном, особенно те, которые касаются отбора сторон и периода сбора данных.

Как уже обсуждалось, выбранное временное окно оказывает влияние, но значения ошибок, найденные Юнгхером, были в порядке, указанном в Тумасьяном. Тем не менее, использование временного окна, заканчивающегося в день выборов (наиболее правдоподобное решение), привело к ошибке в 2,13%, что значительно больше, чем как в первоначальном отчете Тумасьяна, так и в традиционных опросах.

Однако, уже говорилось, что необходимы дальнейшие исследования в отношении периодов сбора данных, и нет никаких оснований предполагать, что подсчет твитов может быть более чувствительным к этому вопросу, чем сентимент-анализ.

Возможно, ключевым аспектом этого метода является отбор кандидатов и партий для мониторинга. Однако, за исключением анализа дела Пиратской партии [81], исследований на этот счет нет: все работы, в которых применялся подсчет твитов, опирались на основные партии.

Отвечая на вопрос, является ли количество твитов для основных партий хорошим показателем намерения проголосовать стоит обратить внимание на следующее. На первый взгляд ответ может показаться неубедительным, половина газет, использующих этот метод, правильно предсказали выборы. Однако, с интуитивной точки зрения, это кажется слишком хорошим, чтобы быть правдой.

В этом отношении весьма поучителен эксперимент, проведенный Гайо-Авелло [101]. Он сравнил производительность как метода подсчета твитов, так и сентимент-анализатора на основе лексики со случайным классификатором.

С этой целью он собрал данные неофициального опроса общественного мнения, проведенного во время президентских выборов в США в Twitter. Веб-сайт под названием TwitVote2 попросил пользователей объявить о своих голосах с помощью твита, помеченного хэштегом #twitvote. Собрав эти твиты, опубликованные в день выборов, он смог найти фактические голоса для ряда пользователей. Кроме того, принимая во внимание долю голосов за обоих кандидатов, можно было бы рассчитать показатели для случайного классификатора и взять их за основу.

Таким образом, он обнаружил, что количество твитов уступает случайному классификатору для обоих кандидатов: незначительно для Обамы и с огромным отрывом для Маккейна. Однако сентимент-классификатор, основанный на лексике, превзошел случайный классификатор для обоих кандидатов.

Другими словами, независимо от того, насколько привлекательно это выглядит, количество необработанных твитов не приближается к случайному выбору, а даже хуже. Следовательно, и несмотря на якобы положительные результаты, такого метода следует избегать в будущем.

#### 3.1.4.2. Подсчет голосов путем сентимент-анализа

Другим популярным методом определения намерений голосования по твитам является сентимент-анализ. Название вводит в заблуждение, поскольку, несмотря на обширные исследования, проведенные в этой области [110], практически все исследования по прогнозированию выборов основывались на простейших методах.

За исключением исследований Бермингема [105] и Джонга [106], которые применили машинное обучение для подготовки своих классификаторов настроений – надо сказать, со смешанными результатами остальные исследования полагались на лексиконы для определения полярности твитов.

О`Коннор [77] был первым, кто использовал этот метод. Он опирался на словарь Уилсона [111], который состоит из списка терминов, помеченных как положительные или отрицательные. Таким образом, твиты могут быть оценены тем или иным способом или даже присвоены оба балла.

Как и в случае с предыдущим методом, этот также привлекателен своей простотой. Однако это также неудовлетворительно. О`Коннор уже находил много примеров неправильно обнаруженных сентиментов, хотя они все еще утверждали: «При довольно большом количестве измерений эти ошибки будут сведены на нет относительно величины, которую мы заинтересованы оценить, совокупного общественного мнения». Это может быть проблематичным, если классификатор выдает разное количество ошибок для каждого кандидата.

В этом отношении эксперимент Гайо-Авелло [101] снова уместен. Он проверил производительность классификатора на основе лексики, аналогичного тому, который использовался О`Коннором обнаружив, что точность для Обамы была довольно высокой (88,8%), но крайне низкой для Маккейна (17,7%). Следовательно, хотя метод является классификатором, он не сбалансирован, и нереалистично ожидать, что ошибки будут устранены при агрегировании результатов.

Наконец, Метаксас [102] провел дополнительные эксперименты с методами, основанными на лексике, подтвердив, во-первых, что их эффективность лишь немного лучше, чем у случайного классификатора, и, во-вторых, что вводящая в заблуждение информация и пропаганда пропускаются или ошибочно интерпретируются как поддержка кандидата. Иначе говоря, методы, основанные на полярности, имеют некоторые недостатки:

1. Упускают из виду тонкости политического языка;
2. Демонстрируют низкую производительность;
3. Получение несбалансированных результатов делает нереалистичным принятие того, что ошибки будут устранены при агрегировании данных.

Таким образом, сентимент-анализ остается открытой задачей в этой области исследований.

### Оценка эффективности

#### Меры оценивания

Последний вопрос, касающийся методов прогнозирования выборов, заключается в том, как их оценивать. Излишне говорить, что необходимы фактические результаты выборов, но что это значит? На каком уровне следует рассчитывать и оценивать прогнозы? Это очевидно для местных выборов, но не так сильно для национальных выборов, которые можно было бы рассчитать, как совокупное всенародное голосование или на более детализированных уровнях.

Текущие исследования дали прогнозы в основном на национальном уровне [78], [81], [104], [105], с несколькими работами, посвященными государственному уровню [101], [102].

Эти прогнозы были оценены по показателям голосов [101], [78], [81], [104], [105], по количеству мест [106], а также как дихотомические решения [102].

При прогнозировании показателей голосования MAE (средняя абсолютная погрешность) обычно используется после исследования Тумасьяна [78]. Этот показатель позволяет исследователям сравнить эффективность их метода с результатами предвыборных опросов.

Прогноз победителя также привлекателен, но он может ввести в заблуждение, поскольку не предоставляется никаких подробностей о том, насколько далеко или близко предсказание было от фактических результатов, и в значительной степени зависит от детализации прогноза. Например, Гайо-Авелло [101] предсказал победу Обамы, которая была правильной, но включала победу в Техасе, которая была неверной.

Более того, результаты Метаксаса [102], приведенные в таблице 2, показывают, что как количество твитов, так и сентимент-анализ позволили правильно угадать половину гонок. Однако, как будет обсуждаться в далее, это не очень хорошая производительность.

Наконец, MAE сильно зависит от каждой гонки и выборов. Например, выборы в Сенат в Кентукки были правильно предсказаны с MAE в 39,6%, в то время как MAE в 6,3% дал неверный прогноз в Калифорнии.

#### Выбор точки отсчета

Следует понимать с каким исходным уровнем сравнивать показатели. Что касается предсказания победителя Метаксас [102] предположил, что: «Учитывая, что исторически действующий кандидат переизбирается примерно 9 раз из 10, базовым показателем для любого компетентного предсказателя должен быть показатель переизбрания действующего кандидата». При таком предположении ясно, что угадывать 50% гонок довольно далеко от компетентности.

Что касается прогнозирования уровня голосования, этот автор не знает о какой-либо базовой линии, подобной предыдущей. Тем не менее, следует использовать результаты непосредственно предшествовавших выборов в качестве прогноза.

Конечно, есть и минусы: например, новые партии, баллотирующиеся на выборах, или коалиции, созданные или распущенные между выборами. Тем не менее, это просто и может дать интуитивный намек на то, насколько «трудными» или «легкими» могут быть предсказания выборов.

Такой базовый уровень был использован для получения данных в таблице 3 и для обсуждения эффективности предыдущих методов при прогнозировании различных выборов. Как таблица, так и обсуждение приведены далее.

|  |  |  |  |
| --- | --- | --- | --- |
| Выборы, г. | Базовая линия, % | Прогнозирование путем подсчета твитов, % | Прогнозирование путем сентимент-анализа, % |
| Выборы президента США, 2008 | 5,86 [101] | 15,87 [101] | 13,10 [101] |
| Федеральные выборы в Германии, 2010 | 3,75 | 1,65 [78];  От 1,51 до 3,34 [81] | Нет результата |
| Выборы в США, 2011 | 8,85 [102] | 17,12 [102] | 7,58 [102] |
| Всеобщие выборы в Ирландии, 2011 | 6,26 | 5,58 [105] | 3,67 [105] |
| Всеобщие выборы в Сингапуре, 2011 | 3,36 | 5,23 [104] | Нет результата |
| Выборы в сенат в Голландии, 2011 | 2,38 | 1,33 [106] | 2 [106] |

Таблица 3 - Производительность, измеренная как MAE для базового прогноза, предсказывающего прошлые показатели голосования, повторения, и двух разных видов прогнозов Twitter

### Возможность предсказания результатов выборов на основе твитов

В таблице II описаны 14 различных попыток предсказать выборы на основе данных Twitter. Только половина из них была успешной. Один из результатов, о котором сообщил Тумасьян [78] – сильно оспаривался Юнгхером [81], а остальные соответствуют прогнозам в статье Метаксаса [102], который смог предсказать половину гонок.

Все это выглядит близким к простой случайности, и заявленные значения MAE напрямую не сопоставимы ни между газетами, ни между гонками на одних и тех же выборах.

В предыдущем разделе была предложена новая базовая линия: предполагая, что прошлые показатели голосования повторятся. Следовательно, любой метод, не отвечающий требованиям такого базового уровня, следует считать неудачным.

Таким образом, показатели эффективности были рассчитаны для каждого из выборов, изученных в литературе:

1. Базового уровня;
2. Прогнозов, основанных, на количестве твитов;
3. Прогнозов, основанных на сентимент-анализе.

К счастью, все статьи, за исключением Джонга [106], непосредственно предоставляют такую информацию. В более позднем случае он был рассчитан на основе результатов, сообщенных авторами (места в Сенате).

В таблице 3 приведены те показатели эффективности, которые могут помочь сделать некоторые выводы о прогностической силе применяемых в настоящее время методов.

#### Достоверность результата, полученного путем подсчета количества твитов

Есть три отчета, в которых прогнозы, основанные на количестве твитов, превосходят исходные данные: отчет Тумасьяна [78], Бермингема [105] и Джонга [106]. Другие три отчета показывают, что такой метод не соответствует базовым показателям: отчеты Гайо-Авелло [101], Метаксаса [102], и Скорика [104].

Кроме того, Юнгхер [81] показал, что:

1. Требуются предварительные важные решения относительно того, какие стороны следует рассмотреть;
2. Эффективность, измеряемая MAE, сильно зависит от используемого временного окна.

Все исследователи, которые воспроизвели метод, упустили незначительные партии – даже те, чьи результаты не соответствовали базовым.

Вкратце, прогноз Twitter, основанный на количестве твитов:

1. Слишком зависит от произвольных решений, таких как партии или кандидаты, подлежащие рассмотрению, или выбор периода для сбора данных;
2. Его производительность слишком нестабильна и сильно зависит от таких настроек;
3. Принимая во внимание представленные результаты в целом, представляется вероятным, что положительные результаты могли быть получены случайно или даже непреднамеренно из-за сбора данных в результате post hoc анализа.

В отсутствие дальнейших исследований, показывающих, что этот метод может последовательно предсказывать будущие результаты для ряда выборов, превосходящих как действующие, так и исходные показатели прошлых результатов, можно сделать вывод, что нет убедительных доказательств того, что это допустимый метод прогнозирования.

#### Достоверность результата, полученного путем сентимент-анализа твитов

Неясно, какое влияние оказывает сентимент-анализ на прогнозы, основанные на Twitter. Исследований, в которых применяется этот метод, меньше, чем тех, которые подсчитывают твиты, и картина, которую они передают, сбивает с толку.

Согласно исследованию Гайо-Авелло [101] сентимент-анализ лучше, чем количество необработанных твитов, но он все еще ниже базового уровня.

Метаксас [102] показывает, что он превосходит не только исходные показатели, но и базовый уровень. Однако он также показал, что методы, основанные на лексике, близки к случайным классификаторам и что доля правильно угаданных результатов не лучше, чем случайность.

Результаты в исследованиях Бермингема [105] и Джонга [106] вносят еще большую путаницу. Первые обнаружили, что сентимент-анализ превосходит базовый, но их метод был усовершенствован, поскольку они включали данные предвыборных опросов, следовательно, были неубедительными. Результаты вторых авторов показывают, что их самый сложный метод превосходит базовый, но не лучше, чем количество необработанных твитов.

Результаты противоречивы. Однако, принимая во внимание, что даже сентимент-анализ, по-видимому, превосходит базовый уровень, ясно, что необходимы дальнейшие исследования в этом направлении.

## Недостатки прогнозирования результатов с помощью Twitter

Предсказательная сила Twitter в отношении выборов была преувеличена, и недавние результаты несопоставимы с результатами оригинальных работ.

Некоторые из сообщенных положительных результатов, вероятно, можно отнести к случайному или непроизвольному извлечению данных, и простые исходные данные во многих случаях обеспечивают лучшую производительность.

Это неудивительно, учитывая, что современные подходы имеют ряд недостатков, которых следует избегать:

1. Все подходы являются результатом post hoc анализа;
2. Производительность следует сравнивать с действительными исходными показателями;
3. Сентимент-анализ применяется с неточностями. Обычно используемые методы немного лучше, чем случайные классификаторы, и не позволяют уловить тонкости политического дискурса;
4. Предполагается, что все твиты заслуживают доверия, хотя это не так;
5. Демографической предвзятостью пренебрегают, даже когда хорошо известно, что социальные сети не являются случайной выборкой населения;
6. Предвзятость самоотбора просто игнорируется. Люди пишут в твиттере на добровольной основе, и, следовательно, данные собираются теми, кто политически активен.

Необходимо выполнить ряд различных задач, прежде чем делать достоверные заявления о прогнозировании выборов на основе данных Twitter:

* Точность анализа настроений политических твитов должна улучшиться. Юмор и сарказм будут играть главную роль;
* Необходимо установить новые показатели надежности: включать твиты, но не ограничиваться этим, обнаруживать пропаганду, дезинформацию и роботов или создать проверку достоверности;
* Требуется фундаментальное исследование демографии Twitter и автоматическое профилирование пользователей по демографическим признакам;
* Необходимы фундаментальные исследования участия пользователей в политике и предвзятости при самостоятельном выборе.

Если не устранить упомянутые недостатки, сосредоточившись на основных направлениях исследований, описанных выше, полученные прогнозы будут иметь сомнительную ценность или вовсе неверны большую часть времени. Это запятнало бы область исследований, которая при надлежащем изучении могла бы быть потенциально полезной.

Следовательно, необходим ряд требований, гарантирующих не только качество результатов, но и общность применяемых методов. Такие требования тесно связаны со схемой определения характеристик, описанной ранее:

1. Эффективность должна оцениваться на основе надежных исходных показателей;
2. Период и метод сбора должны быть четко указаны и обоснованы. Любое предварительное решение относительно кандидатов или партий, подлежащих увольнению, должно быть обоснованным;
3. Чистота данных должна быть гарантирована. То есть должны собираться только твиты пользователей, имеющих право голоса;
4. Вместо упрощенных и грубых подходов следует применять современные методы сентимент-анализа;
5. Лишняя информация в твитах должна быть сведена к минимуму. То есть должны быть предприняты серьезные попытки удалить спам и дезинформацию;
6. Предвзятость в данных должна быть признана и проанализирована. Поощряются попытки устранить демографическую предвзятость. Следует, по крайней мере, проанализировать предвзятость самоотбора.

Однако следует сделать пессимистическое замечание. Все предыдущие соображения являются непременным условием для того, чтобы считать предвыборные прогнозы, основанные на социальных сетях, полностью аналогичными традиционным опросам.

К сожалению, по мере продвижения вниз по списку требований задачи становятся все сложнее и даже недостижимее. Улучшить сентимент-анализ в отношении политических твитов будет сложно, но достижимо. Борьба с демографическим уклоном, по крайней мере, мыслима. Однако учет предвзятости при самоотборе может оказаться нежизнеспособным на общей основе, не используя только данные Twitter.

Следовательно, социальные сети вполне могут дать представление об итогах выборов, и чем лучше методы, тем точнее представление. Однако текущее состояние техники не дает никаких убедительных доказательств в поддержку идеи о том, что оно заменит традиционные опросы в краткосрочной перспективе.