|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Авторы** | **Выборы** | **Период и метод сбора** | **Отбор данных** | **Методы прогнозирования** | **Оценка эффективности** | **Результаты** |
| О`Коннор [77] | Выборы президента США, 2008 | С 02 по 11 месяцы 2008 г.  Использование имен кандидатов в качестве ключевых слов | Отбор не производился | Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Никаких попыток прогнозирования не предпринималось | Корреляция с предвыборными опросами | Существенной корреляции не обнаружено |
| Гайо-Авелло [101] | 01.06.2008 – 03.11.2008 гг.  Использование имен кандидатов в качестве ключевых слов | Твиты с геолокацией на уровне округа.  Попытка удалить данные в соответствии с возрастом пользователя. | Сентимент-анализ на основе лексики;  Индивидуальные голоса;  Агрегированные результаты на государственном уровне;  Показатели голосования | Сопоставление с учетом фактических результатов выборов | MAE 13,10% (неконкурентоспособно с традиционными опросами) |
| Тумасьян [78] | Федеральные выборы в Германии, 2009 | 13.08.2009 – 19.09.2009 гг.  Использование имен кандидатов и партий в качестве ключевых слов | Отбор не производился | Количество твитов;  Агрегированные результаты на национальном уровне;  Показатели голосования | MAE 1,65% (сопоставимо с традиционными опросами) |
| Юнгхер [81] | 13.08.2009 – 27.09.2009 гг.  Использование имен партий в качестве ключевых слов | Нестабильный MAE в зависимости от временного окна, но крупнее MAE, о чем сообщил Тумасьян [78]  Неправильный прогноз при учете всех партий, баллотирующихся на выборах |
| Метаксас [102] | Выборы в США в Массачусетс, 2010 | 13.01.2010 – 20.01.2010 гг.  Использование имен кандидатов в качестве ключевых слов | Количество твитов; Агрегированные результаты на государственном уровне;  Показатели голосования; | Предсказание победителя и MAE | Неправильный прогноз.  MAE 6,3% |
| Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Показатели голосования | Правильный прогноз. MAE 1,2% |
| Выборы в США в Колорадо, 2010 | 26.10.2010 – 01.11.2010 гг.  Имена кандидатов, используемые для фильтрации набора данных «gardenhose» | Количество твитов; Агрегированные результаты на государственном уровне;  Показатели голосования; | Неправильный прогноз.  MAE 24,6% |
| Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Показатели голосования | Правильный прогноз. MAE 12,4% |
| Выборы в США в Неваде, 2010 | Количество твитов; Агрегированные результаты на государственном уровне;  Показатели голосования; | Правильный прогноз. MAE 2,1% |
| Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Показатели голосования | Неправильный прогноз.  MAE 4,7% |
| Выборы в США в Калифорнии, 2010 | Количество твитов; Агрегированные результаты на государственном уровне;  Показатели голосования; | Правильный прогноз. MAE 3,8 % |
| Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Показатели голосования | Неправильный прогноз.  MAE 6,3% |
| Выборы в США в Кентукки, 2010 | Количество твитов; Агрегированные результаты на государственном уровне;  Показатели голосования; | Правильный прогноз. MAE 39,6% |
| Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Показатели голосования | Правильный прогноз. MAE 1,2% |
| Выборы в США в Делавэр, 2010 | Количество твитов; Агрегированные результаты на государственном уровне;  Показатели голосования; | Неправильный прогноз.  MAE 26,5% |
| Сентимент-анализ на основе лексики;  Агрегированные результаты на национальном уровне;  Показатели голосования | Неправильный прогноз.  MAE 19,8% |
| Ливн [103] | Выборы в США, 2010 | 25.03.2007 – 01.11.2010 гг.  Использование твитов и социальных графов для 700 кандидатов | Непригодный.  Этот метод использовал не твиты потенциальных избирателей, а данные кандидатов. | Регрессионные модели для бинарных результатов гонок, которые включали внешние данные;  Агрегированные результаты на государственном уровне;  Прогноз победителя. | Предсказание победителя | Точность 81,5% при использовании только внешних данных.  Точность 83,8% при включении твитов (но не графических данных).  Заметного улучшения нет |
| Бирмингем [105] | Всеобщие выборы в Ирландии, 2011 | 08.02.2011 – 25.02.2011 гг.  Использование имен партий в качестве ключевых слов | Отбор не производился | Количество твитов;  Агрегированные результаты на национальном уровне;  Показатели голосования | Сопоставление с учетом фактических результатов выборов | MAE 5,58% (неконкурентоспособно с традиционными опросами) |
| Сентимент-анализ на основе ML;  Агрегированные результаты на национальном уровне;  Показатели голосования | MAE 3,67% (неконкурентоспособен с традиционными опросами даже после переобучения для использования данных опроса для обучения) |
| Скорик [104] | Всеобщие выборы в Сингапуре, 2011 | 01.04.2011 – 07.05.2011 гг.  Твиты 13 000 сингапурских политически ангажированных пользователей.  Для фильтрации твитов использовались имена партий и кандидатов | Использовались только данные, полученные пользователями, расположенными в Сингапуре. | Количество твитов;  Агрегированные результаты на национальном уровне;  Показатели голосования |  | MAE 5.23%  Безрезультатно, поскольку предвыборные опросы в Сингапуре запрещены |
| Джонг [106] | Выборы в сенат в Голландии, 2011 | 23.02.2011 – 01.03.2011 гг.  Использование имен партий в качестве ключевых слов | Отбор не производился | Количество твитов;  Агрегированные результаты на национальном уровне;  Количество мест в Сенате | Соотношение количества мест по сравнению с фактическими результатами | MAE 1,33%. Конкурирует с традиционными опросами |
| Попытка опровергнуть данные в соответствии с политическими пристрастиями с использованием данных предвыборных опросов | Сентимент-анализ;  Агрегированные результаты на национальном уровне;  Количество мест в Сенате | MAE 2%. Сопоставимо с традиционными опросами |

Таблица 2 - Различные исследования по возможности прогнозирования выборов с использованием данных Twitter характеризуются в соответствии с предложенной выше схемой. Отчеты упорядочиваются в соответствии с датой избрания, а не публикации. Результаты отображаются справа.