Overview

- Problem Motivation
- Related Work
- Problem Statement
- Hardness
- Proposed Solution
- Methodology
- Validation
- Conclusions
Problem Motivation

- There is a vast amount of available geotagged data from social media.
- Is it possible to make accurate predictions about real-life events using this data?
- This can be beneficial for a number of reasons.

Possible applications
- Natural Disasters
- Political Uprising
- Crowdsourcing
- Geo-targeted advertising
Related Work

● Matt Bush, Ivan Lee, and Tony Wu at Stanford University did a study of tweet classification with basic NLP (natural language processing) algorithms
  ○ However, they simply classified tweets into topic (e.g. personal, sports, news, etc.)

● A similar study was conducted at Ohio State University
  ○ In this study, external domain-specific features such as the author's profile data was also used.
Related Work (continued)

- Researchers at the University of Maryland have a project called TwitterStand
  - They tried to create a news aggregation service by processing tweets, as well as looking at popularity of users.
  - Tried to determine which tweets are breaking news.
  - No spatial component.

- Hanan Samet (also at University of Maryland) has a project called NewsStand.
  - This project incorporates a spatial component into the news, but did not use Twitter data.
Contributions and Novelty

- Uses Twitter data and has a spatial component
- Compare to outside data
- Classify the intent of the tweets
  - We want to know if they are pro- or anti-Obama, as opposed to classifying the tweets based on the subject category that they belong to
Problem Statement

- Can we predict how counties will vote in the US Presidential election by classifying the geocoded tweets from those counties?
  - Does the aggregate of tweets mentioning a U.S. presidential candidate's name correctly predict how a county will vote?
    - Why counties?
    - Why tweets?
    - Why U.S. presidential election?
Problem Statement

- **Inputs:** incoming tweets from Twitter stream
- **Outputs:** prediction of which candidate will win each county
- **Constraints**
  - Relatively low proportion of tweets are geotagged (~1%)
  - Many of these incorrectly geotagged
Hardness

● Computational
  ○ High dimensional data
    ■ Each tweet can be thought of as a (sparse) vector of word counts
  ○ Tweets have both a spatial (geolocation) and temporal (timestamp) aspect that can be examined
  ○ Scale

● Spatial
  ○ Difficult to compare point data from Twitter to external geographic data sources
  ○ Difficult to visualize tweet data
    ■ example on next slide
Proposed Solution -- High Level

- Harvest tweet data from important milestones in the election
  - Vice Presidential debate
  - Final Presidential debate
  - Day of the election
- Classify each tweet as pro-Obama, pro-Romney, anti-Obama, anti-Romney, or neither/neutral
- Aggregate results by county
- Validate results based on ground truth data from the election
Methodology

● Obtaining data
  ○ Used tweepy -- Twitter API for Python
  ○ Twitter stream
    ■ searched for political terms: "debate", "Romney", "Obama", "Biden", "Ryan"
      ● Many results contained irrelevant tweets -- limitation of the Twitter stream implementation

● Storing data
  ○ Used SpatiaLite databases
  ○ Wrote a script to put Twitter stream results into database
  ○ Main table in the database contains a record for each tweet, with fields for user, tweet text, location, etc.
Methodology

● Classifying Tweets
  ○ First pass -- determine if tweet is associated with Obama, Romney or neither
    ■ This is done by checking for presence of words associated with each respective candidate.
    ■ Example query:
      ● select lower(tweet) from ___ where lower (tweet) like "%obama%"
  ○ Second pass -- determine if tweet is in favor, in opposition, or neutral to the candidate based on the presence of negative or positive words.

● Aggregating Results
  ○ Spatial join with county polygons from U.S. census
Validation

- Used state and county election results
- Obtained by scraping Politico election results for each state and county
- Compared our predicted election result to the actual election results
State Results

- Correctly predicted 34 of the 48 states correctly -- 70.8333%
- Of the 14 incorrectly predicted states, 9 of those were predicted Obama but voted Romney.
  - Can be due to sampling bias of Twitter.
Tweet percent by county
County Results

- Similar story
- 2158 out of 3109 counties were correctly predicted -- 69.41%
  - majority of incorrectly predicted counties have a very low population
Cartogram of Tweets
Cartogram of Population
Nationwide Popular Vote

- **Predicted**
  - Obama: 52%
  - Romney: 48%

- **Actual**
  - Obama: 51%
  - Romney 48%

- **Prediction is very close to actual results**
  - Note that we did not classify tweets as Gary Johnson, Jill Stein, etc
Conclusions

● Large sampling bias
  ○ Even though we had a lot of tweets, they are more concentrated in specific urban regions
  ○ There are many counties which have no geocoded tweets
  ○ small numbers problem

● Overall very accurate
  ○ tweets correspond with population and election
Future Directions

- Longer period of tweet collection
- More sophisticated algorithms for detecting intent
- Apply same methodology to different events
  - Can even be used for other elections for which there is a lack of available polling data