Mining the Social Web

Asmelash Teka Hadgu
teka@L3S.de

L3S Research Center

April 26, 2016
Outline

INTRODUCTION

USER CLASSIFICATION

NETWORK ANALYSIS

CONTENT ANALYSIS

PRIVACY ISSUES
Introduction

Definition\(^1\)
Web Science is the emergent study of the people and technologies, applications, processes and practices that shape and are shaped by the World Wide Web.

► Studying the online world to understand the offline world.

\(^1\)http://www.websci16.org/
Social web

- Web 2.0 describes World Wide Web sites that emphasize user-generated content, usability, and interoperability.
- The social web is a set of social relations that link people through the World Wide Web.

Figure: Social Media billboard

Twitter

- Politicians use Twitter to mobilize users.
- Companies use Twitter for marketing products.
- Researchers use Twitter to post cfp, share resources and communicate.

*With its open APIs Twitter is a great place for research!*

---

³https://twitter.com/
User classification [1, 2]

How can we automatically construct user profiles?
Applications

- Authoritative users extraction - Discovering expert users for a target topic.
- Personalized web search - Personalized social media posts retrieval.
- User recommendation - Suggesting new interesting users to a target user.
Baseline

- Bag-of-words (BOW) model with weighting (e.g. Tf-idf)
- Each user is represented as a bag of the words in their tweets.
- These words are then used as features to train a classifier.
Feature engineering is key in a machine learning task.

- Profile features
- tweeting behaviour
- linguistic content
- social network features
Profile features

Ideally Profile information should be sufficient. In reality it does not contain enough quality information to be directly used for user classification.
Profile features

Ideally Profile information should be sufficient. In reality it does not contain enough quality information to be directly used for user classification.

- Length of name, Number of alphanumeric chars.
- Capitalization forms in user name
- Use of avatar picture
- Number of followers, friends
- Regular expression matches in bio:
  \( (I|i)(m|am|'m|[0-9]) + (yo|yearold)whiteman|woman|boy|girl \)
Tweeting behaviour features

A set of statistics capturing the way users interact with the micro-blogging service.
Tweeting behaviour features

A set of statistics capturing the way users interact with the micro-blogging service.

- Number of tweets of a user.
- Number and fraction of retweets of a user.
- Ave. number of hashtags and URLs per tweet
- Ave. time and std between tweets
Linguistic content features

Linguistic content contains the user’s lexical usage and the main topics of interest to the user.

\[
\text{proto}(w, c_i) = \left| w, S_i \right| \sum_{n_j=1} \left| w, S_i \right|
\]

where \(\left| w, S_i \right|\) is the number of times the word \(w\) is issued by all users for class \(c_i\).
Linguistic content features

Linguistic content contains the user’s lexical usage and the main topics of interest to the user.

- List of socio-linguistic words such as emoticons, ellipses etc.
- Typical lexical expressions for people in a specific class. prototypical words, hashtags etc. Given $n$ classes, each class $c_i$ containing seed users $S_i$. Each word $w$ from seed users is assigned a score:

$$ proto(w, c_i) = \frac{|w, S_i|}{\sum_{j=1}^{n} |w, S_i|} $$

(1)

where $|w, S_i|$ is the number of times the word $w$ is issued by all users for class $c_i$.

- Latent Dirichlet Allocation (LDA) to generate topics used as features for classification.
- Sentiment words
Social network features

These features contain the social connections between a user and those one follows, replies to or whose messages they retweet. Prototypical ‘friend’ accounts are generated by exploring the social network of users in the training set by bootstrapping as in Eq. (1)

- Number of prototypical friends, percentage number of prototypical friend
- Prototypical replied users, Prototypical retweeted users
Related issues

Manually building *ground truth* data and *validating* result not easy. Usual practices:

- Combine multiple data sources. e.g. scrapping websites that contain users in each class.
- Using key terms in profile of users. Prone to systematic bias and care should be taken on features.
- Using crowdsourcing services such as mechanical turk\(^4\), CrowdFlower\(^5\)

\(^4\)https://www.mturk.com
\(^5\)http://www.crowdflower.com/
Goal: study how social media shape the networked public sphere and facilitate communication between communities with different political orientations.
Framework

- Data gathering
- Identifying political content
- Political communication networks
- Network analysis
Data Set

- Data collected from the Twitter Streaming API\(^6\) between September 14th and November 1st, 2010. Approx. 355 million tweets.
- Data set available at:
  cnets.indiana.edu/groups/nan/truthy

\(^6\)dev.twitter.com/pages/streaming_api
Identifying Political Content

Political communication - any tweet containing at least one *politically relevant hashtag*.

- Political hashtags bootstrapped from seed hashtags (e.g. #p2 and #tcot) using co-occurrence patterns.
- Let $S$ be set of tweets containing seed hashtag and $T$ set of tweets containing another hashtag. Jaccard coefficient between $S$ and $T$:

$$\sigma(S, T) = \frac{|S \cap T|}{|S \cup T|}$$  \hspace{1cm} (2)

if $\sigma$ is big the two hashtags are deemed to be related.

- This process resulted in a corpus of 252,300 (0.07%) politically relevant tweets.
Network data

Political communication networks are generated from the tweets containing politically relevant hashtags where nodes represent users.

▶ In the retweet network an edge runs from a node $A$ to a node $B$ if $B$ retweets content originally broadcast by $A$.

▶ In the mention network an edge runs from $A$ to $B$ if $A$ mentions $B$ in a tweet.
Cluster Analysis

- To establish structure, community detection using a label propagation method for two communities

- Label propagation - Assign an initial arbitrary cluster membership to each node and then iteratively update each node’s label according to the label that is shared by most of its neighbors.

- Modularity to measure segregation. Retweet and mention networks have values of 0.48 and 0.17, respectively.
Finding

The retweet network exhibits two distinct communities of users, while the mention network is dominated by a single massive cluster.

Figure: Political retweet network (left) and mention network (right)
Content Homogeneity

Do clusters have similar content?

- Associate each user with a profile vector of hashtags in their tweets, weighted by frequency.
- Cosine similarity among users.
- Result:

<table>
<thead>
<tr>
<th></th>
<th>Retweet</th>
<th>Mention</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ↔ A</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>B ↔ B</td>
<td>0.20</td>
<td>0.22</td>
</tr>
<tr>
<td>A ↔ B</td>
<td>0.13</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Table: Cosine similarities among user profiles. The average similarities in the retweet and mention networks for pairs of users both in cluster A, both in cluster B, and for users in different clusters.
Political Polarization

Do clusters in the retweet network correspond to groups of users of similar political alignment?

▶ Qualitative content analysis from social science.
▶ One author annotates 1,000 random users as ‘left’ or ‘right’.
▶ Another user annotates 200 random users from the 1,000 users above.
▶ Inter annotator agreement measured using Cohen’s Kappa

\[ k = \frac{P(\alpha) - P(\epsilon)}{1 - P(\epsilon)} \]

where \( P(\alpha) \) is observed rate of agreement between annotators and \( P(\epsilon) \) is expected rate of random agreement given relative frequency of each class label.
Content Analysis[4]

Political Hashtag Trends, PHT, an analysis tool for political polarization of Twitter hashtags in the US.

[Image: POLITICAL TWITTER TRENDS from YAHOO LABS]

Trending political hashtags grouped by weekly leaning, from Mar. 25, 2012 to Mar. 31, 2012

Ordered by left leaning
1. #occupycongress
2. #yes4m
3. #paul
4. #budget4all
5. #nom

Ordered by right leaning
1. #tpn
2. #nyhbl
3. #addthis
4. #romneycaresucks
5. #usmc

7http://politicalhashtagtrends.sandbox.yahoo.com/
Data Set

- Start with a set of *seed political users* such as @BarackObama and @MittRomney whose political leaning is known.
- Get their tweets using Twitter REST API\(^8\)

\(^8\)https://dev.twitter.com/docs/api/1.1
Collect users that retweet seed users’ tweets.
Filtering Users by Location

- Named entity recognition to limit analysis to the U.S.

Yahoo! Placemaker™ Beta
Evaluating Data Quality

- Dataset validation against Web Directories.

- Precision = 0.98, 0.93 for Wefollow and Twellow respectively.

- Manual inspection: “greatest environmentalist. Also, despise republicans”
Detecting Political Hashtags

- Look into co-occurrence with seed political hashtags (#p2, #tcot, #gop, #ows) and (‘obama’, ‘romney’, ‘politic’, ‘liberal’, ‘conservative’, ‘democ’, or ‘republic’)
- Volume filtering to avoid rare hashtags.
Computing Trending Score

- trending - currently popular. Having a higher volume than expected.

\[
\text{trend}(h, w) := \frac{f(h, w) / \sum_{h' \in H} f(h', w)}{\sum_{u \leq w} f(h, u) / \sum_{h' \in H} \sum_{u \leq w} f(h', u)}
\]

Examples:

- #obamagotosama: 01 May 2011 to 08 May 2011.
- Non-trending hashtags: #vote, #democracy.
Assigning a Leaning to Hashtags

Using Voting approach: Let $v_L$ denote the aggregated user volume of $h$ in $w$ for the left leaning. Let $V_L$ denote the total left user volume of all hashtags in $w$. Similarly for $v_R$ and $V_R$.

$$\text{Lean}(h,w) := \frac{v_L}{V_L} + \frac{2}{V_L + V_R}$$

where a leaning of 1.0 is fully left and 0.0 is fully right.
Privacy Issues

Who Wants to Get Fired?[5]

**Figure**: FireMe! on the news
FireMe! Homepage

Figure: Fireme! Home page

http://fireme.l3s.uni-hannover.de/fireme.php
Problem Definition

Study posting behaviour of individuals who tweet they hate their job or boss.
Observation

- Many users are not aware of their audience.
- Build an alerting system to address the danger of public online data.

Collect ‘lovers’ - Users posting positive messages like: ‘I love my job’, ‘my boss is the best’.

Crawled users for one week, June 18 - June 26, 2012. 21,851 haters and 44,710 lovers.

Randomly select 10,000 users from each group. Get the latest 200 tweets of each user.
Characterising groups:

- lovers are more connected. Three times as many followers, and 20% more friends.
- haters are more active in terms of tweeting speed - twice as many tweets per day.
- haters link to their facebook profile in their bio.
FireMe! Features

Figure: FireMe! alert page

- Shows user, her hate tweet and FireMeter! score
- FireMeter! score fraction of hate tweets in the latest 100 tweets of a user.
Alert Impact

- Sent out over 4000 alert messages to haters. ‘What are you going to do about it?’
- Feedback: 42% delete hate tweet, 18% change privacy setting, 40% don’t care.
- Overall 60% users concerned about their personal data.
FireMe! Features...

Figure: FireMe! provides a leaderboard
Validation and more analysis . . .

- Majority of users deleted the compromising tweets.
- Tone of alert messages (neutral, aggressive, action) is important.
Thank you! Any Questions?


