Can Social Media tell us something about our lives?

Vasileios Lampos

Computer Science Department
University of Sheffield

March, 2013
Outline

- Motivation, Aims [Facts, Questions]
- Data
  - Nowcasting Events
  - Extracting Mood Patterns
  - TrendMiner – Extracting Political Opinion
- Conclusions
Facts

We started to work on those ideas back in 2008, when...

- **Web** contained *1 trillion* unique pages (Google)

- **Social Networks** were rising, e.g.
  - *Facebook*: 100m (2008) → *>1 billion* active users (October, 2012)
  - *Twitter*: 6m (2008) → *500m* active users (July, 2012)

- **User behaviour** was changing
  - Socialising via the Web
  - Giving up privacy  *(Debatin et al., 2009)*
Some general questions

• Does user generated text posted on Social Web platforms include useful information?

• How can we extract this useful information... 
  ... automatically? Therefore, not we, but a machine.

• Practical / real-life applications?

• Can those large samples of human input assist studies in other scientific fields?
  Social Sciences, Psychology, Epidemiology
Why Twitter?

• Has a lot of content that is **publicly accessible**
• Provides a well-documented **API** for several types of data collection
• **Opinions** and **personal statements** on various domains
• Connection with current affairs (usually in **real-time**)
• Some content is **geo-located**
• Option for **personalised modelling**
• **... and we got good results from the very first, simple experiment!**
The Data (2/3)

What does a @tweet look like?

Figure 1: Some biased and anonymised examples of tweets (limit of 140 characters/tweet, # denotes a topic)

(a) (user will remain anonymous)
(b) they live around us

another demo covered by citizens today in Thessaloniki int'l fair. Citizen journalism on a speed rise in #Greece. check #deth and #rbnews

(c) citizen journalism

i think i have the flu but i still look fabulous

(d) flu attitude
Data Collection & Preprocessing

• The easiest part of the process...
  ◦ not true! → Storage space, crawler implementation, parallel data processing, new technologies (e.g., Map-Reduce) (Preotiuc et al., 2012)

• Data collected via Twitter’s Search API:
  ◦ collective sampling
  ◦ tweets geo-located in 54 urban centres in the UK
  ◦ periodical crawling (every 3 or 5 minutes per urban centre)

• Data collected via Twitter’s REST API:
  ◦ user-centric sampling
  ◦ preprocessing to approximate user’s location (city & country)
  ◦ ... or manual user selection from domain experts
  ◦ get their latest tweets (3,000 or more)

• Several forms of ground truth (flu/rainfall rates, polls)
Nowcasting Events from the Social Web
‘Nowcasting’?

We do not predict the future, but infer the present — $\delta$

i.e. the very recent past

Figure 2: Nowcasting the magnitude of an event ($\varepsilon$) emerging in the real world from Web information

Our case studies: nowcasting (a) flu rates & (b) rainfall rates (?!)

V. Lampos bill@lampos.net
What do we get in the end?

This is a **regression** problem (**text regression** in NLP)

\[ \forall \text{ time interval } i \text{ we aim to infer } y_i \in \mathbb{R} \text{ using text input } x_i \in \mathbb{R}^n \]

![Graph showing inferred rainfall rates for Bristol, UK (October, 2009)](image)

**Figure 3:** Inferred **rainfall rates** for Bristol, UK (October, 2009)
Candidate features \((n\text{-grams})\): \(\mathcal{C} = \{c_i\}\)

Set of **Twitter posts** for a time interval \(u\): \(\mathcal{P}(u) = \{p_j\}\)

Frequency of \(c_i\) in \(p_j\):

\[
g(c_i, p_j) = \begin{cases} 
\varphi & \text{if } c_i \in p_j, \\
0 & \text{otherwise.}
\end{cases}
\]

– Boolean, maximum value for \(\varphi\) is \(1\) –

**Score** of \(c_i\) in \(\mathcal{P}(u)\):

\[
s(c_i, \mathcal{P}(u)) = \frac{\sum_{j=1}^{\mathcal{P}(u)} g(c_i, p_j)}{|\mathcal{P}(u)|}
\]
Set of **time intervals**: $\mathcal{U} = \{ u_k \} \sim 1 \text{ hour, 1 day, ...}$

**Time series** of candidate features **scores**:

$$X^{(\mathcal{U})} = \left[ x^{(u_1)} \ldots x^{(u_{|\mathcal{U}|})} \right]^T,$$

where

$$x^{(u_i)} = \left[ s \left( c_1, \mathcal{P}(u_i) \right) \ldots s \left( c_{|\mathcal{C}|}, \mathcal{P}(u_i) \right) \right]^T$$

**Target variable** (event):

$$y^{(\mathcal{U})} = \left[ y_1 \ldots y_{|\mathcal{U}|} \right]^T$$
Solve the following optimisation problem:

$$\min_w \| X^{(U)} w - y^{(U)} \|_2^2$$

s.t.  \( \| w \|_1 \leq t \),

\[ t = \alpha \cdot \| w_{OLS} \|_1, \quad \alpha \in (0, 1]. \]

• Least Absolute Shrinkage and Selection Operator (LASSO)

\[ \arg\min_w \| X^{(U)} w - y^{(U)} \|_2^2 + \lambda \| w \|_1 \]

(Tibshirani, 1996)

• Expect a sparse \( w \) (feature selection)

• Least Angle Regression (LARS) – computes entire regularisation path (\( w \)'s for different values of \( \lambda \)) (Efron et al., 2004)
LASSO is **model-inconsistent**:  
- inferred sparsity pattern may deviate from the true model, e.g., when predictors are highly correlated (Zhao and Yu, 2006)

- bootstrap LASSO (Bolasso) performs a more robust feature selection (Bach, 2008):  
  - in each bootstrap, input space is sampled with replacement  
  - apply LASSO (LARS) to select features  
  - select features with nonzero weights in all bootstraps

- better alternative — **soft-Bolasso**:  
  - a less strict feature selection  
  - select features with nonzero weights in $p\%$ of bootstraps  
  - (learn $p$ using a separate validation set)

- **weights** of selected features determined via OLS regression
Methodology (5/5) — Simplified summary

**Observations:** \( X \in \mathbb{R}^{m \times n} \) (\( m \) time intervals, \( n \) features)

**Response variable:** \( y \in \mathbb{R}^m \)

**For** \( i = 1 \) to **number of bootstraps**

- Form \( X_i \subset X \) by sampling \( X \) with replacement
- Solve LASSO for \( X_i \) and \( y \), i.e. learn \( w_i \in \mathbb{R}^n \)
- Get the \( k \leq n \) features with nonzero weights

**End_For**

Select the \( v \leq n \) features with nonzero weight in \( p \% \) of the bootstraps

Learn their weights with OLS regression on \( X^{(v)} \in \mathbb{R}^{m \times v} \) and \( y \)
How do we form candidate features?

- Commonly formed by indexing the **entire corpus**
  
  \[ (Manning, Raghavan and Schütze, 2008) \]

- We extract them from Wikipedia, Google Search results, Public Authority websites (e.g., NHS)

**Why?**

- reduce **dimensionality** to bound the error of LASSO

\[
\mathcal{L}(w) \leq \mathcal{L}(\hat{w}) + Q, \text{ with } Q \sim \min \left\{ \frac{W_1^2}{N}, \frac{p}{N}, \frac{W_1^2}{N} + \frac{W_1}{\sqrt{N}} \right\} 
\]

  \( p \) candidate features, \( N \) samples, empirical loss \( \mathcal{L}(\hat{w}) \) and \( \|\hat{w}\|_1 \leq W_1 \) \hfill (Bartlett, Mendelson and Neeman, 2011)

- **Harry Potter Effect!**
**Figure 4:** Events co-occurring (correlated) with the inference target may affect feature selection, especially when the sample size is small.

![Graph showing event scores over day numbers with legend: Flu (England & Wales), Hypothetical Event I, Hypothetical Event II.](Lampos, 2012a)
The ‘Harry Potter’ effect (2/2)

Table 1: Top 1-grams correlated with flu rates in England/Wales (06–12/2009)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>latitud</td>
<td>Latitude Festival</td>
<td>0.9367</td>
</tr>
<tr>
<td>flu</td>
<td>Flu epidemic</td>
<td>0.9344</td>
</tr>
<tr>
<td>swine</td>
<td></td>
<td>0.9212</td>
</tr>
<tr>
<td>harri</td>
<td>Harry Potter Movie</td>
<td>0.9112</td>
</tr>
<tr>
<td>slytherin</td>
<td></td>
<td>0.9094</td>
</tr>
<tr>
<td>potter</td>
<td></td>
<td>0.8972</td>
</tr>
<tr>
<td>benicassim</td>
<td>Benicàssim Festival</td>
<td>0.8966</td>
</tr>
<tr>
<td>graduat</td>
<td>Graduation (?)</td>
<td>0.8965</td>
</tr>
<tr>
<td>dumbledore</td>
<td>Harry Potter Movie</td>
<td>0.8870</td>
</tr>
<tr>
<td>hogwart</td>
<td></td>
<td>0.8852</td>
</tr>
<tr>
<td>quarantin</td>
<td>Flu epidemic</td>
<td>0.8822</td>
</tr>
<tr>
<td>gryffindor</td>
<td>Harry Potter Movie</td>
<td>0.8813</td>
</tr>
<tr>
<td>ravenclaw</td>
<td></td>
<td>0.8738</td>
</tr>
<tr>
<td>princ</td>
<td></td>
<td>0.8635</td>
</tr>
<tr>
<td>swineflu</td>
<td>Flu epidemic</td>
<td>0.8633</td>
</tr>
<tr>
<td>ginni</td>
<td>Harry Potter Movie</td>
<td>0.8620</td>
</tr>
<tr>
<td>weaslei</td>
<td></td>
<td>0.8581</td>
</tr>
<tr>
<td>hermion</td>
<td></td>
<td>0.8540</td>
</tr>
<tr>
<td>draco</td>
<td></td>
<td>0.8533</td>
</tr>
</tbody>
</table>

Solution: ground truth with some degree of variability

(Lampos, 2012a)
About n-grams

1-grams

• decent (dense) representation in the Twitter corpus
• unclear semantic interpretation
  Example: “I am not sick. But I don’t feel great either!”

2-grams

• very sparse representation in tweets
• sometimes clearer semantic interpretation

Experimental process indicated that...

a hybrid combination* of 1-grams and 2-grams
delivers the best inference performance

* refer to (Lampos, 2012a)
Figure 5: Font size is proportional to the weight of each feature; flipped n-grams are negatively weighted. All words are stemmed (Porter, 1980).

(Lampos and Cristianini, 2012)
Figure 6: Font size is proportional to the weight of each feature; flipped n-grams are negatively weighted. All words are stemmed (Porter, 1980).

(Lampos and Cristianini, 2012)
Examples of inferences

(a) Central England/Wales (flu)

(b) South England (flu)

(c) Bristol (rain)

Figure 7: Examples of flu and rainfall rates inferences from Twitter content (Lampos and Cristianini, 2012)
## Performance figures

**Table 2:** RMSE for flu rates inference (5-fold cross validation), 50m tweets, 21/06/2009–19/04/2010

<table>
<thead>
<tr>
<th>Method</th>
<th>1-grams</th>
<th>2-grams</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td>12.44±2.37</td>
<td>13.81±3.29</td>
<td>11.62±1.58</td>
</tr>
<tr>
<td>Bolasso</td>
<td>11.14±2.35</td>
<td>12.64±2.57</td>
<td>10.57±2.2</td>
</tr>
<tr>
<td>CART ensemble**</td>
<td>9.63±5.21</td>
<td>13.13±4.72</td>
<td>9.4±4.21</td>
</tr>
</tbody>
</table>

* As implemented in (Ginsberg et al., 2009)

**Table 3:** RMSE (in mm) for rainfall rates inference (6-fold cross validation), 8.5m tweets, 01/07/2009–30/06/2010

<table>
<thead>
<tr>
<th>Method</th>
<th>1-grams</th>
<th>2-grams</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td>2.91±0.6</td>
<td>3.1±0.57</td>
<td>4.39±2.99</td>
</tr>
<tr>
<td>Bolasso</td>
<td>2.73±0.65</td>
<td>2.95±0.55</td>
<td>2.60±0.68</td>
</tr>
<tr>
<td>CART ensemble**</td>
<td>2.71±0.69</td>
<td>2.72±0.72</td>
<td>2.64±0.63</td>
</tr>
</tbody>
</table>

** Classification and Regression Tree (Breiman et al., 1984) & (Sutton, 2005)**
Figure 8: Flu Detector uses the content of Twitter to nowcast flu rates in several UK regions

(Lampos, De Bie and Cristianini, 2010)
Extracting Mood Patterns from the Social Web
Computing a mood score

Table 4: Mood terms from WordNet Affect

<table>
<thead>
<tr>
<th>Fear</th>
<th>Sadness</th>
<th>Joy</th>
<th>Anger</th>
</tr>
</thead>
<tbody>
<tr>
<td>afraid</td>
<td>depressed</td>
<td>admire</td>
<td>angry</td>
</tr>
<tr>
<td>fearful</td>
<td>discouraged</td>
<td>cheerful</td>
<td>despise</td>
</tr>
<tr>
<td>frighten</td>
<td>disheartened</td>
<td>enjoy</td>
<td>enviously</td>
</tr>
<tr>
<td>horrible</td>
<td>dysphoria</td>
<td>enthusiastic</td>
<td>harassed</td>
</tr>
<tr>
<td>panic</td>
<td>gloomy</td>
<td>exciting</td>
<td>irritated</td>
</tr>
<tr>
<td>...</td>
<td>(92 terms)</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Mood score computation for a time interval $d$ using $n$ mood terms

$$ms_d = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i(t_d)}{N(t_d)}$$

$c_i(t_d)$: count of term $i$ in the Twitter corpus of day $d$

$N(t_d)$: number of tweets for day $d$

Using the sample of $d$ days, compute a standardised mood score:

$$ms_{d}^{std} = \frac{ms_d - \mu_{ms}}{\sigma_{ms}}$$
The mood of the nation (1/5)

**Figure 9**: Daily time series (actual & their 14-point moving average) for the mood of Joy based on Twitter content geo-located in the UK

**933 Day Time Series for Joy in Twitter Content**

![Graph showing daily time series for joy in Twitter content](image)

- Normalised Emotional Valence
- Date
- Raw joy signal
- 14-day smoothed joy
- Key dates and events marked:
  - XMAS
  - Valentine's Day
  - Easter
  - Halloween
- Dates: Jul 09 to Jan 12

*(Lansdall, Lampos and Cristianini, 2012a&b)*
Can Social Media tell us something about our lives?

The mood of the nation (2/5)

Figure 10: Daily time series (actual & their 14-point moving average) for the mood of Anger based on Twitter content geo-located in the UK

(Lansdall, Lampos and Cristianini, 2012a&b)
The mood of the nation (3/5)

Window of 100 days: 50 before & after the point of interest

\[ ms_i^{\text{std}} = \mu \left( ms_{i+1 \rightarrow i+50}^{\text{std}} \right) - \mu \left( ms_{i-50 \rightarrow i-1}^{\text{std}} \right) \]

Figure 11: Change point detection using a 100-day moving window

(Lansdall, Lampos and Cristianini, 2012a)
The mood of the nation (4/5)

Figure 12: Projections of 4-dimensional mood score signals (joy, sadness, anger and fear) on their top-2 principal components (PCA) – Twitter content from 2011

(a) Days of the week (2011)

(b) Days of the year (2011)

Cluster I
New Year (1), Valentine’s (45), Christmas Eve (358), New Year’s Eve (365)

Cluster II
O.B. Laden’s death (122), Winehouse’s death + Breivik (204), UK riots (221)

(Lampos, 2012a)
Figure 13: Mood of the Nation uses the content of Twitter to nowcast mood rates in several UK regions

(Lampos, 2012a)
Circadian mood patterns (1/3)

Compute **24-h mood score patterns**

**Mood score** computation for a **time interval** $u = 24$hours using $n$ **mood terms** (WordNet) and a sample of $D$ **days**:

\[
M_s(u) = \frac{1}{|D|} \sum_{j=1}^{D} \left( \frac{1}{n} \sum_{i=1}^{n} s_{f_i}^{(t_j,u)} \right)
\]

\[
s_{f_i}^{(t_d,u)} = \frac{f_i^{(t_d,u)} - \bar{f}_i}{\sigma_{f_i}}, \quad i \in \{1, \ldots, n\}.
\]

$f_i^{(t_d,u)}$: normalised frequency of a mood term $i$ during time interval $u$ in day $d \in D$
Figure 14: Circadian (24-hour) mood patterns based on UK Twitter content
Circadian mood patterns (3/3)

Figure 15: Autocorrelation of circadian mood patterns based on hourly lags revealing daily and weekly periodicities

(a) Fear

(b) Sadness

(c) Joy

(d) Anger

... further analysis on those patterns (in collab. with domain experts) under submission
TrendMiner Project

Extracting political opinion from Social Media
A few words about the project...

- **TrendMiner** is an EU-FP7 project

- Several participants incl. the Univ. of Sheffield & Southampton (UK) and DFKI (Germany)

- Aims to form **methods for interpreting** the vast stream of **online information**

- Our focus on analysis of Twitter content → **political opinion, financial indicators**

- Work in progress and under submission process → **cannot** go into much detail!
Some new challenges

• Aim: **model voting intention**
  ○ regression task
  ○ multiple outputs

• Overcome **limitations** of previous methods
  ○ use of sentiment analysis taxonomies → language specific, restrictive
  ○ **combined modelling** of word frequencies and the domain of users?
  ○ **multi-task learning** → exploit correlations in the feature space
  ○ multi-task & **multi-domain** learning
    → model political opinion + financial indicators jointly

• Proper **evaluation**
  ○ *k*-fold cross-validation may sometimes be misleading
  ○ can we actually predict future values?
A snapshot of the results

\[ \mathbf{v}_i = \mathbf{u}^T \mathbf{X} \mathbf{w} + \beta \]

(plus multi-task learning)

**Figure 16**: 50 voting intention polls (YouGov) and their respective inferred values for the Conservative (RMSE: 1.78%), Labour (1.59%) and Liberal Democrat (1.05%) parties (Nov. 2011 to Feb. 2012)

(a) Voting intention polls

(b) Voting intention inferences
Qualitative evaluation is also essential...

• Some domains may be represented by **smooth** trends (e.g., political domain)

• Predictions could be easy in that context
  → how do we know we are not **overfitting**?

• Perform qualitative analysis using the selected features (words, users and tweets)
  ○ Do the selected words and users make some sense?
  ○ Does their combination make sense? → score single tweets

• Possibly better models when increasing the statistical evidence (multi-task learning)
Conclusions – Did they tell us anything?

- Social Media hold valuable information

- We can develop methods to extract portions of this information automatically
  - detect, quantify, nowcast events
  - extract collective mood patterns
  - model other domains (such as politics)

- User generated input + other features → tell/reveal something about the users & their context

- Side effect: what about our privacy? ...
In collaboration with...

Prof. Nello Cristianini, University of Bristol (Ph.D. Advisor)
Prof. Ricardo Araya, University of Bristol (Psychiatry)
Dr. Tijl De Bie, University of Bristol
Thomas Lansdall-Welfare, University of Bristol

Dr. Trevor Cohn, University of Sheffield (TrendMiner)
Daniel Preotiuc-Pietro, University of Sheffield (TrendMiner)
The end.

Any questions?

Download the slides from http://www.lampos.net/research/presentations-and-posters
References


