Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks

Thibaut THONE T
Guillaume CABANAC
Mohand BOUGHANEM
Karen PINEL-SAUVAGNAT

IRIT, Université de Toulouse, CNRS

7 November 2017
Motivation

- Massive amount of opinions on the Web $\Rightarrow$ Need for automated methods to identify, classify and summarize opinions

- Traditional opinion mining research mainly focused on product/service review analysis $\Rightarrow$ Identification of a review’s polarity w.r.t. a target: positive/negative

Images and reviews taken from Wikipedia and Amazon.com, February 2016.
Motivation

... But need to go beyond plain positive/negative opinions ⟷ viewpoint-based opinions

E.g., to deal with filter bubbles [Pariser, 2011] & echo chambers [Sunstein, 2009]
Task

How to mitigate filter bubbles & echo chambers? ⇒ Build **unbiased viewpoint summaries**
This work is the **first step**: discover **viewpoints** and **topics** from social networking data.
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**.
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**.

**Text content** component
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted *text content* and *social interactions*

**Text content** component

- Observed data: **tokens** occurring in *documents* posted by **users** → 3 nested plates
- Latent **topics** assigned to each token
- Latent **viewpoints** assigned at document-level
SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**.

**Text content** component

Following the **Topic-Aspect Model** from [Paul+, AAAI ’10], definition of **four word types** specified by switch variables $\ell$ (level) and $x$ (route):

- **Background** words
  \[ \ell = 0, \; x = 0 \]

- **Viewpoint** words
  \[ \ell = 0, \; x = 1 \]

- **Topic** words
  \[ \ell = 1, \; x = 0 \]

- **Viewpoint-topic** words
  \[ \ell = 1, \; x = 1 \]
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted text content and social interactions.

**Social interaction component**
SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted text content and social interactions.

**Social interaction component**

**Outgoing interactions** for user $u = $ interactions initiated by $u$ on another user (recipient $r$)

Following **SN-LDA** from [Sachan+., WSDM '14], viewpoints assigned to outgoing interactions (**homophily**)
We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted **text content** and **social interactions**.

Social interaction component

But outgoing interactions **insufficient** for some users

We propose to also exploit **incoming interactions**
SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted text content and social interactions.

**Social interaction** component

**Incoming interactions** for user $u =$ interactions initiated by another user (**sender** $s$) on $u$

Viewpoint assigned to the **document** being interacted upon
SNVDM: The Social Network Viewpoint Discovery Model

We defined the **Social Network Viewpoint Discovery Model** to jointly discover topics and viewpoints from posted text content and social interactions.

**Posterior inference**

Approximate inference based on **Collapsed Gibbs Sampling**

- Dirichlet/Bernoulli distributions $\sigma$, $\psi$, $\theta$, $\pi$, $\phi$, $\xi$ **integrated out**
- Successively **sample** discrete latent variables $\ell$, $x$, $z$, $v$, $v'$ from their posterior distributions (i.e., given observations $w$, $s$, $r$)
Limits of SNVDM’s Social Interaction Component

Some users have **very few social interactions**

⇒ Difficult to identify their viewpoints based on scarce **direct** interactions
Limits of SNVDM’s Social Interaction Component

We propose to extend SNVDM to leverage “aquaintances of acquaintances” (≈ friends of friends)

How? \[\implies \text{Generalized Pólya Urn scheme}\]
Using **Generalized Pólya Urn** in SNVDM requires minor changes in collapsed Gibbs sampling.

E.g., for outgoing interaction $o$ from user $u$ on user $u'$:

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V} \cdot n_{vu'} + \mu \frac{1}{U}}{n_{u*} + \eta} \cdot \frac{n_{vu'} + \mu}{n_{v*} + \mu}$$

**SNVDM vs ...**

---

**...SNVDM-GPU**

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V} \cdot \sum_{u''=1}^{U} A_{u''u'} n_{vu''} + \mu \frac{1}{U}}{n_{u*} + \eta} \cdot \frac{\sum_{u''=1}^{U} A_{u''u} n_{vu''} + \mu}{\sum_{u''=1}^{U} A_{u''u} n_{vu''} + \mu}$$

---

**SNVDM-GPU**
Experimental Setup: Datasets & Evaluated Models

- **Twitter** datasets from [Brigadir+, WebSci ’15] on the **2014 Scottish Independence Referendum** ($v = \text{Yes/No}$) and the **2014 US Midterm Elections** ($v = \text{Democrat/Republican}$)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Users</th>
<th>#Tweets</th>
<th>#Tokens</th>
<th>Vocabulary</th>
<th>#Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indyref</td>
<td>589</td>
<td>575</td>
<td>270,075</td>
<td>2,043,204</td>
<td>38,942</td>
</tr>
<tr>
<td>Midterms</td>
<td>767</td>
<td>778</td>
<td>113,545</td>
<td>975,199</td>
<td>25,312</td>
</tr>
</tbody>
</table>


**Experimental Setup: Datasets & Evaluated Models**

- **Twitter** datasets from [Brigadir+, WebSci ’15] on the **2014 Scottish Independence Referendum** \( (v = \text{Yes/No}) \) and the **2014 US Midterm Elections** \( (v = \text{Democrat/Republican}) \)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Users</th>
<th>#Tweets</th>
<th>#Tokens</th>
<th>Vocabulary</th>
<th>#Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indyref</td>
<td>589</td>
<td>575</td>
<td>270,075</td>
<td>2,043,204</td>
<td>38,942</td>
</tr>
<tr>
<td>Midterms</td>
<td>767</td>
<td>778</td>
<td>113,545</td>
<td>975,199</td>
<td>25,312</td>
</tr>
</tbody>
</table>

**State-of-the-art baselines:**

- **Topic-Aspect Model** (TAM) from [Paul+, AAAI ’10]
  - \( \Rightarrow \) **Only text content**
    to discover viewpoints and topics

- **Social Network Latent Dirichlet Allocation** (SN-LDA) from [Sachan+, WSDM ’14]
  - \( \Rightarrow \) **Text content** and outgoing interactions
    to discover **communities** \( (\approx \text{viewpoints}) \) and topics

- **Viewpoint and Opinion Discovery Unification Model** (VODUM) from [Thonet+, ECIR ’16]
  - \( \Rightarrow \) **Text content**
    to discover viewpoints and topics, and **parts of speech**
    to distinguish between topic words and viewpoint-topic words
Experimental Setup: Datasets & Evaluated Models

Twitter datasets from [Brigadir+, WebSci ’15] on the 2014 Scottish Independence Referendum ($v = \text{Yes/No}$) and the 2014 US Midterm Elections ($v = \text{Democrat/Republican}$)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Users</th>
<th>#Tweets</th>
<th>#Tokens</th>
<th>Vocabulary</th>
<th>#Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indyref</td>
<td>589</td>
<td>575</td>
<td>270,075</td>
<td>2,043,204</td>
<td>38,942</td>
</tr>
<tr>
<td>Midterms</td>
<td>767</td>
<td>778</td>
<td>113,545</td>
<td>975,199</td>
<td>25,312</td>
</tr>
</tbody>
</table>

State-of-the-art baselines:

- **Topic-Aspect Model** (TAM) from [Paul+, AAAI ’10]
- **Social Network Latent Dirichlet Allocation** (SN-LDA) from [Sachan+, WSDM ’14]
- **Viewpoint and Opinion Discovery Unification Model** (VODUM) from [Thonet+, ECIR ’16]

Proposed models:

- **SNVDM**
- **SNVDM-GPU ($\tau = 10$)**: only 10 most interacting acquaintances used in Generalized Pólya Urns
- **SNVDM-GPU ($\tau = \infty$)**: all acquaintances used in Generalized Pólya Urns
Evaluation: Viewpoint Clustering

Clustering of users’ viewpoints on Indyref in terms of Purity and NMI (error bars = 95% CI)

Observation 1: consistent results across different numbers of topics
Evaluation: Viewpoint Clustering

Clustering of users’ viewpoints on Indyref in terms of Purity and NMI (error bars = 95% CI)

Observation 2: SNVDM, SNVDM-GPU ($\tau = 10$), SNVDM-GPU ($\tau = \infty$) > all baselines
**Evaluation: Viewpoint Clustering**

Clustering of users’ viewpoints on Indyref in terms of Purity and NMI (error bars = 95% CI)

Observation 3: \( \text{SNVDM-GPU (} \tau = \infty \text{)} > \text{SNVDM-GPU (} \tau = 10 \text{)} > \text{SNVDM} \implies \text{GPU beneficial} \)
Evaluation: Viewpoint Clustering

Clustering of users’ viewpoints on Midterms in terms of Purity and NMI (error bars = 95% CI)

Observation 4: similar trends on Midterms but greater improvement for our models over baselines
Evaluation: Impact of Social Network Sparsity

Clustering of users’ viewpoints on Indyref for different degrees of network sparsity ($T = 10$)

Observation: performance degraded for lower percentage of interactions
Evaluation: Qualitative Analysis

Most probable **topic words** and **viewpoint-topic words** for topics from Indyref and Midterms

<table>
<thead>
<tr>
<th>Neutral</th>
<th>Viewpoint: Yes</th>
<th>Viewpoint: No</th>
</tr>
</thead>
<tbody>
<tr>
<td>#indyref</td>
<td>#voteyes</td>
<td>#indyref</td>
</tr>
<tr>
<td>scotland</td>
<td>yes</td>
<td>uk</td>
</tr>
<tr>
<td>independence</td>
<td>scotland</td>
<td>salmond</td>
</tr>
<tr>
<td>vote</td>
<td>independence</td>
<td>#bettertogether</td>
</tr>
<tr>
<td>campaign</td>
<td>westminster</td>
<td>#scotdecides</td>
</tr>
<tr>
<td>scottish</td>
<td>vote</td>
<td>separation</td>
</tr>
<tr>
<td>uk</td>
<td>independent</td>
<td>currency</td>
</tr>
<tr>
<td>people</td>
<td>country</td>
<td>thanks</td>
</tr>
<tr>
<td>future</td>
<td>#yes</td>
<td>today</td>
</tr>
<tr>
<td>independent</td>
<td>#scotland</td>
<td>say</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>energy</td>
<td>#actonclimate</td>
<td>#4jobs</td>
</tr>
<tr>
<td>house</td>
<td>climate</td>
<td>#obamacare</td>
</tr>
<tr>
<td>new</td>
<td>#p2</td>
<td>#jobs</td>
</tr>
<tr>
<td>gas</td>
<td>change</td>
<td>gop</td>
</tr>
<tr>
<td>natural</td>
<td>#climatechange</td>
<td>obama</td>
</tr>
<tr>
<td>#energy</td>
<td>clean</td>
<td>bills</td>
</tr>
<tr>
<td>#ff</td>
<td>oil</td>
<td>jobs</td>
</tr>
<tr>
<td>#kxl</td>
<td>energy</td>
<td>house</td>
</tr>
<tr>
<td>support</td>
<td>#gop</td>
<td>act</td>
</tr>
<tr>
<td>economic</td>
<td>seec</td>
<td>watch</td>
</tr>
</tbody>
</table>

- Reasonable **coherence** of topic words and viewpoint-topic words
- Topic words indeed **unbiased** towards any viewpoints
- Use of viewpoint-specific **hashtags** and mention of different issues for different viewpoints
Conclusion and Research Directions

- **SNVDM(-GPU)**: models to jointly discover viewpoints and topics in social networks, leveraging both **posted text content** and **social interactions**

  Take-home message: **social interactions** are key for viewpoint discovery in social networks!

- What’s next?
  - Integrate **time dimension** and **geolocation**, e.g., to analyze candidate support during elections
  - Design a **viewpoint summarization** framework to provide Internet users with **more diversified content** and thus mitigate the filter bubble and echo chamber phenomenon
Acknowledgments & Questions

Thanks **SIGIR, SIGWEB** and the **US NSF** for providing travel grants!

Questions?

@tthonet
thonet@irit.fr

@gcabanac
cabanac@irit.fr

@MohBoughanem
boughanem@irit.fr

@karenatw
sauvagnat@irit.fr

Code and data available at: [https://github.com/tthonet/SNVDM](https://github.com/tthonet/SNVDM)
References


Appendix: Baseline SNVDM-WII

Ablated version of SNVDM: **SNVDM-WII** (without incoming interactions)
### Appendix: Execution Time

Execution time (in seconds) of one Gibbs sampling iteration on Indyref (with $T = 10$) and Midterms (with $T = 15$)

<table>
<thead>
<tr>
<th>Method</th>
<th>Indyref</th>
<th>Midterms</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAM</td>
<td>1.45</td>
<td>0.87</td>
</tr>
<tr>
<td>SN-LDA</td>
<td>1.18</td>
<td>0.64</td>
</tr>
<tr>
<td>VODUM</td>
<td>2.78</td>
<td>1.85</td>
</tr>
<tr>
<td>SNVDM-WII</td>
<td>2.08</td>
<td>1.08</td>
</tr>
<tr>
<td>SNVDM</td>
<td>2.49</td>
<td>1.15</td>
</tr>
<tr>
<td>SNVDM-GPU ($\tau = 10$)</td>
<td>3.47</td>
<td>1.34</td>
</tr>
<tr>
<td>SNVDM-GPU ($\tau = \infty$)</td>
<td>14.67</td>
<td>2.56</td>
</tr>
</tbody>
</table>
The compound **Dirichlet-Multinomial** distribution (used in LDA-based topic models) can be interpreted as an **urn sampling metaphor** with an **over-replacement** policy.

1. Randomly draw a ball from the urn
2. Duplicate the drawn ball
3. Put back in the urn the original ball and its duplicate

---

**Diagram:**
- **Urn**
- **Infinite ball generator**

---

**Appendix: Simple Pólya Urn Scheme**
The Simple Pólya Urn scheme can be generalized by modifying the replacement rule to exploit similarities between balls’ colors [Mahmoud, 2008]

1. Randomly draw a ball from the urn
2. Duplicate the drawn ball and generate parts of balls for those "similar" to the drawn ball
3. Put back in the urn the original ball, its duplicate and the parts of similar balls

Infinite ball generator

\[ A(\square, \square) = 1 \]
\[ A(\square, \bigcirc) = 1/2 \]
\[ A(\bigcirc, \bigcirc) = 1/2 \]
Using **Generalized Pólya Urn** in SNVDM requires minor changes in collapsed Gibbs sampling. E.g., for outgoing interaction $o$ from user $u$ on user $u'$:

$$p(v'_{uo} = v | r_{uo} = u', \text{rest}) \propto \frac{n_{uv} + \eta \frac{1}{V}}{n_u + \eta} \cdot \frac{n_{vu'} + \mu \frac{1}{U}}{n_v + \mu}$$

**SNVDM vs ...**

The **addition matrix** $A$ defines the **weight** to put on count $n_{vu''}$ for each $u''$:

$$A_{u'u''} = \begin{cases} 
1 & \text{if } u' = u'', \\
\lambda & \text{if } u'' \text{ is among top } \tau \text{ acquaintances of } u', \\
0 & \text{otherwise}
\end{cases}$$

with $0 \leq \lambda \leq 1$ ($\lambda = 0 \implies \text{“vanilla” SNVDM}$) and $\tau \in \mathbb{N}$