Representativeness-aware Aspect Analysis for Brand Monitoring in Social Media

Lizi Liao, Xiangnan He, Zhaochun Ren, Liqiang Nie, Huan Xu, Tat-Seng Chua
National University of Singapore

Motivation

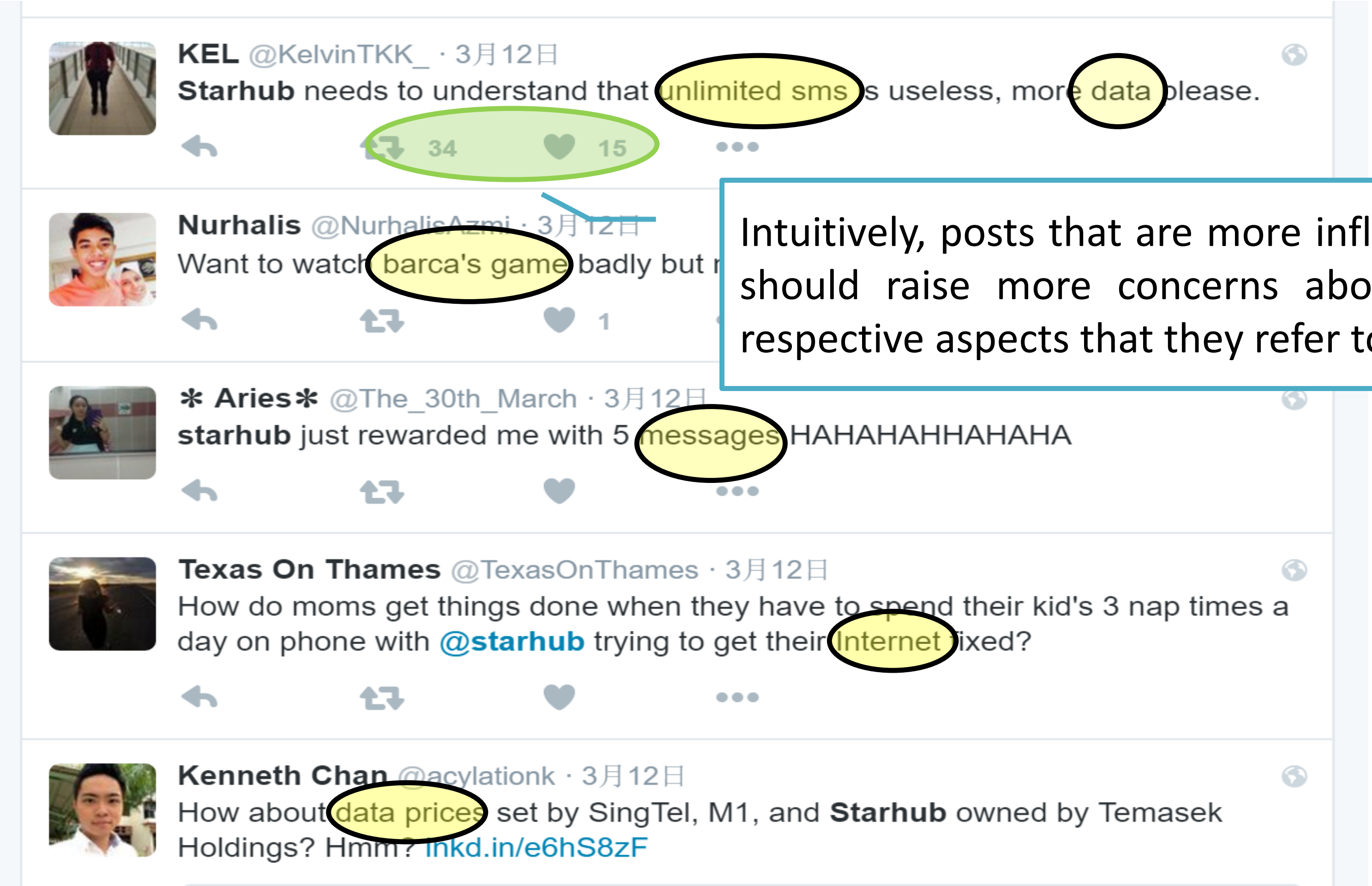
- Fast responding nature
- Success of social media

monitoring the reputation of brands and the opinions of general public

Often referred to as social media listening, brand monitoring is essential for companies to build harmonious relationships with customers and protect their reputation.

-- Glance et al., 2005, Haruechaiyasak et al., 2013

Examples



KEL @KelvinTKK_ · 3月12日

Starhub needs to understand that unlimited sms is useless, more data please.

34 15

Nurhalis @NurhalisAzmi · 3月12日

Want to watch barca's game badly but r

1

✻ Aries✻ @The_30th_March · 3月12日

starhub just rewarded me with 5 messages HAHAHAHHAHAHA

Texas On Thames @TexasOnThames · 3月12日

How do moms get things done when they have to spend their kid's 3 nap times a day on phone with **@starhub** trying to get their Internet fixed?

Kenneth Chan @acylationk · 3月12日

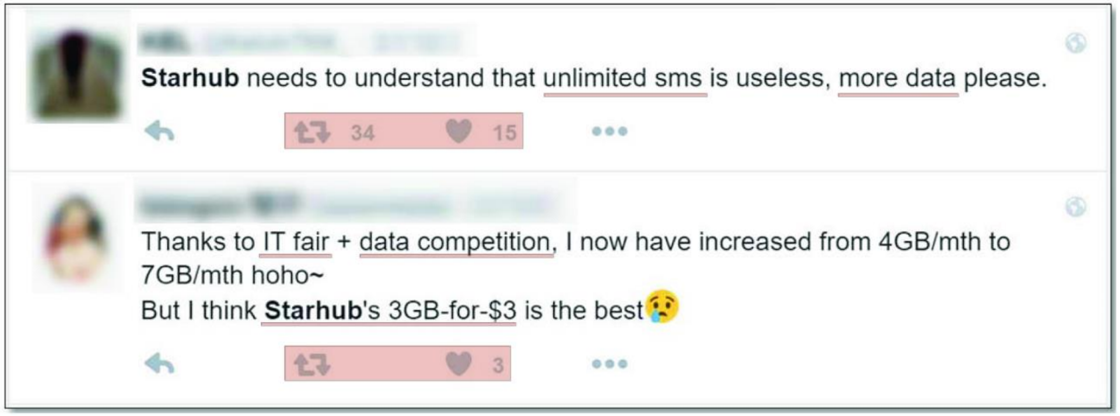
How about data prices set by SingTel, M1, and **Starhub** owned by Temasek Holdings? Hmm? lnkd.in/e6hS8zF

Intuitively, posts that are more influential should raise more concerns about the respective aspects that they refer to.

Users talk about some aspects and posts absorb different attention

Challenges

If a company can automatically identify the representative aspects from the fast-evolving social media data, it can perform fine-grained aspect-level analysis and react to customers’ response timely



[redacted]

Starhub needs to understand that unlimited sms is useless, more data please.

34 15

[redacted]

Thanks to IT fair + data competition, I now have increased from 4GB/mth to 7GB/mth hoho~
But I think Starhub's 3GB-for-$3 is the best😢

3

- To distinguish such users’ preference from texts → non-trivial for machines to automatically figure it out.
- One of the key challenges lies in identifying the representative aspects which not only cover detailed information but also represent customers’ intent.
- In addition to identifying salient aspects, it is beneficial to also select posts that are most representative of the data collection

Previous Works

- Aspect extraction
  - The first type of work treats an aspect as a term, usually a noun or noun phrase that describes the specific properties of products.
    - early approaches extract aspects by considering the term frequency and leveraging dependency relation rules [Hu and Liu, 2004; Popescu and Etzioni, 2005]
    - other approaches model it as a sequence labelling task, applying hidden markov model and conditional random field for aspect identification [Kobayashi et al., 2007; Jakob and Gurevych, 2010]
  - The second type of work treats an aspect as a group of terms.
    - utilize statistical topic models to identify aspects as term distributions [Moghaddam and Ester, 2012], [Paul and Girju,2010]

We opt for a middle way. We extract noun or noun phrases in posts as aspect candidates and incorporate the relation between aspect candidates into the definition of representiveness.

Previous Works

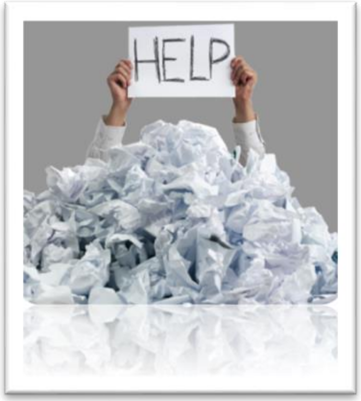
- Extractive summarization
  - One of the standard methods is Maximum Marginal Relevance (MMR) [Carbonell and Goldstein, 1998].
  - [Long et al., 2009] defines the best summary as the one that has the minimum information distance to the entire document set
  - [Lin and Bilmes, 2010] uses non-monotone submodular set functions to perform extractive summarization

It is worth noting that the existing aspect analysis and summarization works have mainly focused on documents. To perform aspect analysis on the social media data, it is crucial to account for the varying impact of posts.

Special Requirements

- First, the algorithm should have the flexibility to be easily guided towards specific aspects.
  - companies may change their focus of monitoring dynamically. For example, after launching a new product, the company would like to know customers' response to it.

- Second, the algorithm needs to be efficient.
  - to provide brand monitoring service online, there is often a need to sift through a huge amount of data and respond quickly.



Definition of Representativeness

- Impact of posts
  - τ_j denotes the impact of p_j . Denoting the number of retweets and likes as r_j and l_j , we set

$$\tau_j = \left(\frac{1 + \alpha r_j + \beta l_j}{1 + \alpha r_{max} + \beta l_{max}} \right)^{\eta}$$

- A post’s representativeness score for an aspect
  - *exact term matching scheme is insufficient (TF-IDF)*

$$\mathcal{R}_i(p_j) = \tau_j \frac{\sum_{a_s \in \mathcal{A}_j} fsim(a_s, a_i)}{|\mathcal{A}_j|}$$

For example, if a salient aspect “data usage” occurs in a post (while “data plan” does not), we cannot say the post is not representative for “data plan”.

-- Character-level edit distances to find the match of tokens, spelling errors or abbrev

-- Apply WordNet similarity weighted token-level edit distance of aspect candidates to calculate fuzzy similarity score

[Chaudhuri et al., 2003] Robust and efficient fuzzy match for online data cleaning

Definition of Representativeness

- R score of a posts set X for an aspect i

$$\mathcal{R}_i(\mathcal{X}) = 1 - \prod_{p_j \in \mathcal{X}} (1 - \mathcal{R}_i(p_j))$$

		$\mathcal{R}_{\mathcal{X}}(\text{WiFi})$
🟦	omg im so done with singtel's wifi, so slow	0.9
🟥	Singtel.... your wifi...... I cannot la I rly cannot	0.6

$$\mathcal{R}_{\text{🟦🟥}}(\text{WiFi}) = 1 - \text{P(neither 🟦 nor 🟥 } \mathcal{R}\ \text{WiFi)}$$

$$= 1 - (1 - 0.9)\ (1 - 0.6) \ = \ 0.96$$

$$\boxed{\mathcal{R}_{\text{🟦}}(\text{WiFi}) \ < \ \underbrace{0.96 \ < \ \mathcal{R}_{\text{🟦}}(\text{WiFi}) + \mathcal{R}_{\text{🟥}}(\text{WiFi})}_{\textbf{Diminishing returns}}}$$

Definition of Representativeness

- The representativeness of a posts set X for the whole posts collection P, which is defined as

$$\mathcal{R}(\mathcal{X}) = \sum_{a_i \in \mathcal{A}} w_i \mathcal{R}_i(\mathcal{X})$$

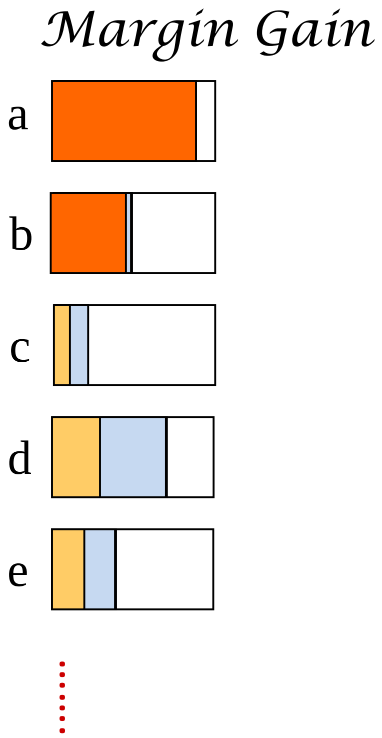
- Given the posts collection P and a budget k, our task is to find k posts that maximizes the representativeness for the whole collection.

$$\mathcal{X}^* = \underset{\mathcal{X} \subseteq \mathcal{P}:|\mathcal{X}|=k}{\arg\max} \sum_{a_i \in \mathcal{A}} w_i \mathcal{R}_i(\mathcal{X})$$

Submodular function

Simple Greedy

Greedy algorithm

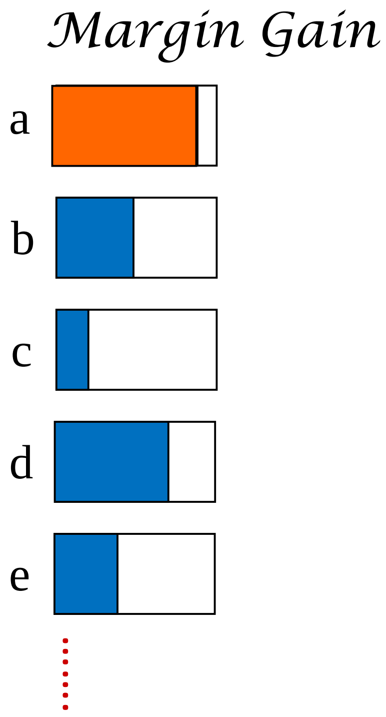


$$MargGain_p = \mathcal{R}(\mathcal{X} \cup \{p\}) - \mathcal{R}(\mathcal{X}).$$

- A simple greedy can obtain good results
  - at least *1-1/e* of optimal [Nemhauser et al ’78]
- But
  - Greedy algorithm is slow
    - scales as *O*(|P|K*)*

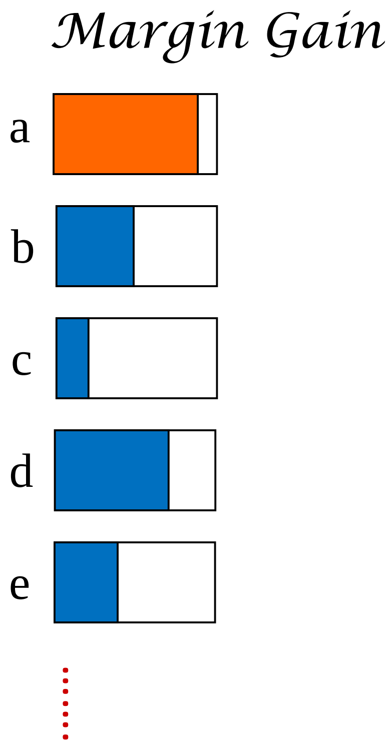
Speed up

- Fast greedy algorithm:
  - Keep an ordered list of marginal gain m_i from previous iteration
  - Re-evaluate m_i only for top post
  - Re-sort and prune



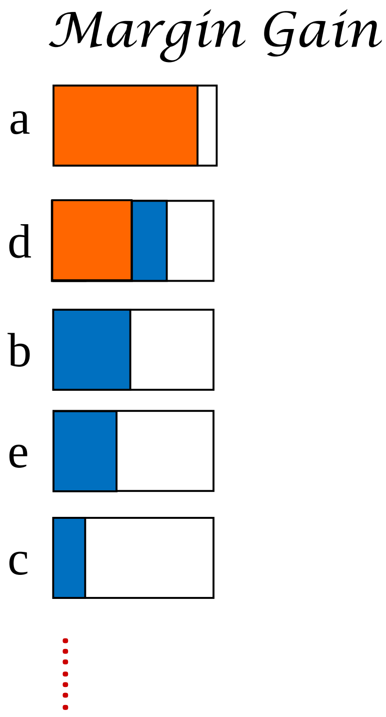
Speed up

- Fast greedy algorithm:
  - Keep an ordered list of marginal gain m_i from previous iteration
  - Re-evaluate m_i only for top post
  - Re-sort and prune



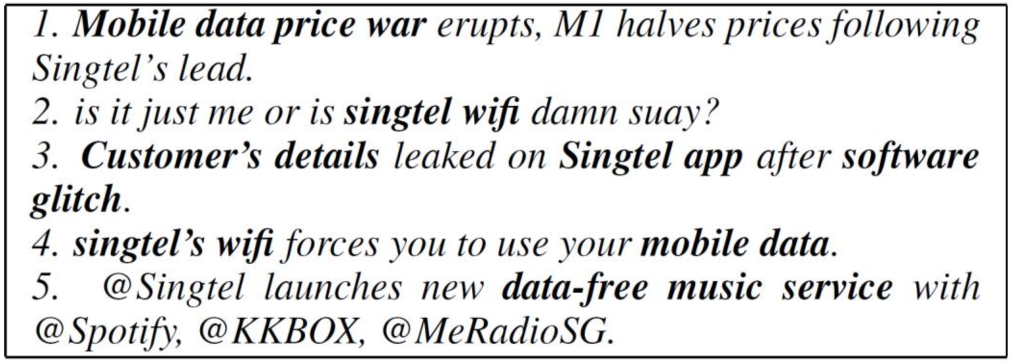
Speed up

- Fast greedy algorithm:
  - Keep an ordered list of marginal gain m_i from previous iteration
  - Re-evaluate m_i only for top post
  - Re-sort and prune



Interesting observation

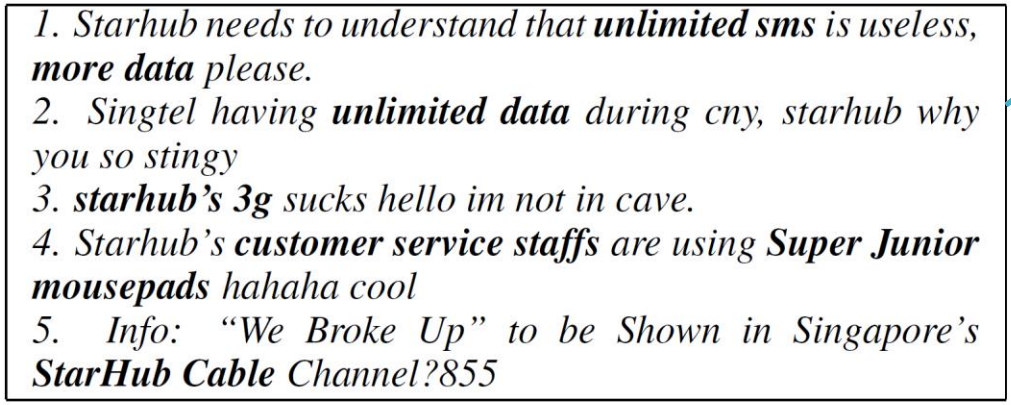
Table 2: Top five posts selected for the Singtel dataset.



*1. **Mobile data price war** erupts, M1 halves prices following Singtel's lead.*
*2. is it just me or is **singtel wifi** damn suay?*
*3. **Customer's details** leaked on **Singtel app** after **software glitch**.*
*4. **singtel's wifi** forces you to use your **mobile data**.*
*5. @Singtel launches new **data-free music service** with @Spotify, @KKBOX, @MeRadioSG.*

To obtain a more comprehensive sense of customers' view and to get better sense of whether a commercial strategy works or not via social media, it might be useful to monitor competing brands at the same time.

Table 3: Top five posts selected for the StarHub dataset.



*1. Starhub needs to understand that **unlimited sms** is useless, **more data** please.*
*2. Singtel having **unlimited data** during cny, starhub why you so stingy*
*3. **starhub's 3g** sucks hello im not in cave.*
*4. Starhub's **customer service staffs** are using **Super Junior mousepads** hahaha cool*
*5. Info: "We Broke Up" to be Shown in Singapore's **StarHub Cable** Channel?855*

the top posts show that users complain a lot about data usage

NExT++ NUS-Tsinghua Centre for Extreme Search

A Joint Research Collaboration Between NUS & Tsinghua University

THANK YOU

NExT research is supported by the National Research Foundation,
Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.