Representativeness-aware Aspect Analysis for Brand Monitoring in Social Media*

Lizi Liao[1], Xiangnan He[1], Zhaochun Ren[2], Liqiang Nie[3], Huan Xu[4], Tat-Seng Chua[1]
[1]School of Computing, National University of Singapore
[2]Data Science Lab, JD.com
[3]School of Computer Science and Technology, Shandong University
[4]ISyE, Georgia Institute of Technology
{liaolizi.llz, xiangnanhe, nieliqiang}@gmail.com, renzhaochun@jd.com,
huan.xu@isye.gatech.edu, dcscts@nus.edu.sg

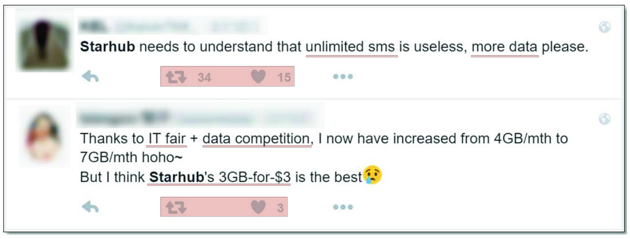
Abstract

Owing to the fast-responding nature and extreme success of social media, many companies resort to social media sites for monitoring the reputation of their brands and the opinions of general public. To help companies monitor their brands, in this work, we delve into the task of extracting representative aspects and posts from users' free-text posts in social media. Previous efforts treat it as a traditional information extraction task, and forgo the specific properties of social media, such as the possible noise in user generated posts and the varying impacts; In contrast, we extract aspects by maximizing their *representativeness*, which is a new notion defined by us that accounts for both the coverage of aspects and the impact of posts. We formalize it as a submodular optimization problem, and develop a FastPAS algorithm to jointly select representative posts and aspects. The FastPAS algorithm optimizes parameters in a greedy way, which is highly efficient and can reach a good solution with theoretical guarantees. Extensive experiments on two datasets demonstrate that our method outperforms state-of-the-art aspect extraction and summarization methods in identifying representative aspects.

1 Introduction

Nowadays, more and more people express their opinions on social media sites like Facebook and Twitter. The fast-responding and broad diffusion nature of those social media sites bring about an increasing demand for companies to monitor the reputation of their brands and the reception by general public. Recent studies by [Glance *et al.*, 2005; Haruechaiyasak *et al.*, 2013] have verified that brand monitoring over social media streams can help companies maintain harmonious relationships with customers and protect their reputations.

Studies related to brand monitoring have been carried out at the document, sentence and aspect levels. Aspect-level analysis is often desired as it provides more fine-grained information about a brand. Figure 1 shows an illustrative example of users' tweets about StarHub, a telecommunication company. Looking into the tweets, we can see that the first user prefers "more data" over "unlimited sms"; here we denote such noun or noun phrases as aspect candidates. In addition, the relatively large number of retweets and likes reveals that the post has been well recognized by other users. This provides an important evidence for StarHub to identify customer needs and improve their services. While it seems to be easy for human to distinguish such users' preference from texts, it is 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 [Liao *et al.*, 2016]. In this case, data usage is more appropriate than sms and other less indicative aspects. If a company can automatically identify the representative aspects from the fast-evolving social media data, the company can perform fine-grained aspect-level analysis and react to customers' response timely [Zhang *et al.*, 2016a].



[redacted]

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

[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😢

Figure 1: An example of users' tweets about StarHub.

Previous studies on aspect identification have treated it as a traditional text analysis task, which either developed some language rules to extract featured terms as aspects [Hu and Liu, 2004; Liu *et al.*, 2015] or grouped terms into aspects with statistical topic models [Paul and Girju, 2010; Ren *et al.*, 2016]. These methods, however, usually work on well-structured data such as the customer reviews and do not consider the varying impacts of posts. Intuitively, posts that are more influential (*e.g.,* voted by many users) should raise more concerns about the respective aspects that they refer to. In addition to identifying salient aspects, it is beneficial to also select posts that are most representative of the data collection, which can assist companies in making decisions. This is related to another line of previous research, i.e. extractive summarization; which aims to select important sentences

*The corresponding author is Xiangnan He.

from documents. However, existing efforts have mainly focused on documents instead of short posts in social media.

Targeting at brand monitoring, we highlight two requirements other than simply extracting a set of representative aspects and posts. First, 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. Thus, the algorithm should have the flexibility to be easily guided towards specific aspects. Second, to provide brand monitoring service online, there is often a need to sift through a huge amount of data and respond quickly [He *et al.*, 2016]. Thus, the algorithm needs to be efficient. In this paper, we tackle the above challenges by developing a new method to automatically select representative aspects and posts from social media. We first define a new notion of *representativeness* of aspect, which can help companies re-adjust their focus of brand monitoring dynamically. Next, we formulate the task as a submodular optimization problem, and design an efficient greedy algorithm named FastPAS (short for *Fast Posts and Aspects Selection*), which can compute a near-optimal solution with theoretical guarantees. Extensive experiments on two Twitter datasets show that our method outperforms existing information extraction methods. In addition, our results suggest that monitoring the competitors' brands might also benefit one's own brand monitoring in social media.

2 Related Work

Existing studies on aspect identification (or aspect extraction) can be categorized into two types. The first type of work treats an aspect as a term, usually a noun or noun phrase that describes the specific properties of products. For example, early approaches [Hu and Liu, 2004; Popescu and Etzioni, 2005] extract aspects by considering the term frequency and leveraging dependency relation rules; other approaches [Kobayashi *et al.*, 2007; Jakob and Gurevych, 2010] model it as a sequence labeling task, applying hidden markov model and conditional random field for aspect identification. The second type of work treats an aspect as a group of terms [Moghaddam and Ester, 2012], for example, [Paul and Girju, 2010] utilize statistical topic models to identify aspects as term distributions. In this work, 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*. By doing this, our method manages to provide easily explainable services to end users (a merit of the first type) [He *et al.*, 2015] as well as offer higher granularity aspects (a merit of the second type).

As we address the problem by jointly selecting representative posts and aspects, it is also related to the extractive summarization task [He *et al.*, 2012]. One of the standard methods is *Maximum Marginal Relevance* (MMR) [Carbonell and Goldstein, 1998]. A major problem of MMR is that the decision made is only based on the current iteration, which can be suboptimal. In contrast, our method has a more solid constant-factor approximation guarantee. Along this line, [Long *et al.*, 2009] defines the best summary as the one that has the minimum information distance to the entire document set, and [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 [Liao *et al.*, 2014], it is crucial to account for the varying impact of posts (which is evidenced by the explicit surrogate like the number of likes and retweets). Moreover, our method is specifically tailored for the downstream application of brand monitoring, being more flexible and efficient than existing methods.

3 Problem Definition

We first introduce the notion of *representativeness*, highlighting the key diminishing return property of our design. Then we formulate the task as a submodular optimization problem that selects representative posts and aspects simultaneously.

3.1 Definition of Representativeness

Given a set of posts $\mathcal{P} = \{p_1, \cdots, p_N\}$, we first extract aspect candidates (which are nouns or noun phrases). We slightly abuse the use of "aspect" and "aspect candidate" in case of no ambiguity. Let the set of aspect candidates be $\mathcal{A} = \{a_1, \cdots, a_M\}$, then we present each post p_j as a tuple $< \mathcal{A}_j, \tau_j >$, where $\mathcal{A}_j$ denotes the set of aspect candidates the post contains; and τ_j denotes the impact of p_j. Denoting the number of retweets and likes for p_j as r_j and l_j, we set

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

where α and β adjust the importance of retweets and likes, respectively. We found that $\eta = 0.75$ returns a modest improvement over the linear version with $\eta = 1$.

Now, we define a post's representativeness score for an aspect candidate. An intuitive way is to weight the aspect using the traditional term weighting schemes, such as TF-IDF. However, we argue that the exact term matching scheme is insufficient for our aspect candidates. 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". With this in mind, we define each post p_j's representiveness $\mathcal{R}_i(p_j)$ for an aspect candidate a_i, i.e.,

$$\mathcal{R}_i(p_j) = \tau_j \, \frac{\sum_{a_s \in \mathcal{A}_j} fsim(a_s, a_i)}{|\mathcal{A}_j|}, \quad (2)$$

where $fsim(\cdot, \cdot) \in [0, 1]$ denotes the similarity score between two aspects, which is calculated from fuzzy matching [Chaudhuri *et al.*, 2003]: we first use character-level edit distances to find the match of tokens, which captures spelling errors or abbreviations. Then we apply WordNet similarity weighted token-level edit distance of aspect candidates to calculate fuzzy similarity score. The representativeness score regarding the aspect a_i is estimated by the average similarity to aspects in p_j. In addition, the post's impact τ_j is also incorporated, so that influential posts are more likely to be representative for its aspects. The denominator $|\mathcal{A}_j|$ punishes long posts containing many aspects.

Next, we define the representativeness score of a posts set $\mathcal{X}$ for an aspect a_i, to pave the way for the task of post selection. Intuitively, when the aspect a_i does not occur in the post set $\mathcal{X}$, adding a new post that covers a_i will gain the most; further adding posts that contain a_i should be rewarded less and less. This diminishing return property is known as *submodularity* in discrete optimization, which is a discrete analog of convexity [Lovász, 1983]. It motivates us to design the function as follows,

$$\mathcal{R}_i(\mathcal{X}) = 1 - \prod_{p_j \in \mathcal{X}} (1 - \mathcal{R}_i(p_j)). \tag{3}$$

This design helps to select posts with fewer redundant aspects. We theoretically prove the submodularity property of $\mathcal{R}_i(\mathcal{X})$ next, which is crucial to our efficient algorithm for posts and aspects selection.

Proof. Equation 3 defines a non-decreasing function such that for any post set $\mathcal{X} \subseteq \mathcal{Y}$, we will have $\mathcal{R}_i(\mathcal{X}) \leq \mathcal{R}_i(\mathcal{Y})$[1]. It states that if we add a new post to a small set $\mathcal{X}$, the reward is no smaller than adding the post to the larger set $\mathcal{Y}$, as formulated in the following theorem.

Theorem 1. *For post sets $\mathcal{X} \subseteq \mathcal{Y}$ and post $p \notin \mathcal{Y}$, it holds that,*

$$\mathcal{R}_i(\mathcal{X} \cup \{p\}) - \mathcal{R}_i(\mathcal{X}) \geq \mathcal{R}_i(\mathcal{Y} \cup \{p\}) - \mathcal{R}_i(\mathcal{Y}).$$

Proof. We first consider the reward of adding post p to the set $\mathcal{Y}$,

$$\mathcal{R}_i(\mathcal{Y} \cup \{p\}) - \mathcal{R}_i(\mathcal{Y}) = \prod_{p_j \in \mathcal{Y}} (1 - \mathcal{R}_i(p_j)) \cdot \mathcal{R}_i(p).$$

Since $\mathcal{R}_i(p_j) \in [0, 1]$, we have $(1 - \mathcal{R}_i(p_j)) \in [0, 1]$. Considering the fact that $\mathcal{X} \subseteq \mathcal{Y}$, we can get,

$$\begin{aligned} & \prod_{p_j \in \mathcal{Y}} (1 - \mathcal{R}_i(p_j)) \cdot \mathcal{R}_i(p) \\ \leq & \prod_{p_j \in \mathcal{X}} (1 - \mathcal{R}_i(p_j)) \cdot \mathcal{R}_i(p) = \mathcal{R}_i(\mathcal{X} \cup \{p\}) - \mathcal{R}_i(\mathcal{X}). \end{aligned}$$

Clearly, we can see that $\mathcal{R}_i(\mathcal{X})$ is submodular. □

3.2 Problem Formulation

In real-world application of brand monitoring, companies usually hope to extract only a small set of posts that contain influential and representative aspects for further analysis. Motivated by this scenario, we consider the representativeness of a posts set $\mathcal{X}$ for the whole posts collection $\mathcal{P}$, which is defined as,

$$\mathcal{R}(\mathcal{X}) = \sum_{a_i \in \mathcal{A}} w_i \mathcal{R}_i(\mathcal{X}), \tag{4}$$

where $w_i \in \mathcal{W}$ denoting the non-negative weight of the aspect a_i, which can be pre-defined by the end user, otherwise uniformly set as 1. It serves as a hyper-parameter that can be flexibly tuned to adjust the aspect focus for selecting representative posts. For example, a telecommunication company is more interested with the aspect "data plan" than "camera resolution", and a news company may have more focus on trending topics. To put more focus on a certain aspect, one just need to set the corresponding w_i to a relatively large number. Moreover, by further introducing an exponential decay on the weight for aspect terms captured in the posts selection procedure, we further avoids redundancy. We will detail this later in Section 4.2. Note that submodularity is closed under nonnegative linear combinations. Therefore, $\mathcal{R}(\mathcal{X})$ also holds the diminishing return property.

In order to find the k posts which are most representative for the whole posts collection, we formulate the problem as follows. Given the posts collection $\mathcal{P}$ and a budget k, our task is to find k posts that maximizes the *representativeness* for the whole collection. Mathematically, the optimization problem is formalized as,

$$\mathcal{X}^* = \operatorname*{arg\,max}_{\mathcal{X} \subseteq \mathcal{P} : |\mathcal{X}| = k} \sum_{a_i \in \mathcal{A}} w_i \mathcal{R}_i(\mathcal{X}). \tag{5}$$

4 Optimization Algorithms

After formulating our problem in Equation 5, we now aim to find an efficient algorithm to solve it. In this section, we first present a greedy optimization solution for the problem; we then leverage its monotone property to speed it up, making it suitable to run on large-scale datasets.

4.1 Greedy Optimization Solution (PAS)

It is known that submodular functions can be minimized in polynomial time; however, maximizing a submodular function is an NP-complete problem [Feige *et al.*, 2011]. It is computationally expensive to get the exact optimal solution. Fortunately, the classic result of [Nemhauser *et al.*, 1978] shows that by applying a simple greedy algorithm to solve Equation 5, we can obtain a $(1 - \frac{1}{e})$ lower bound approximation of the optimal solution. Next, we describe our greedy strategy for solving Equation 5.

We start from $\mathcal{X} = \phi$, which is an empty set. We then iteratively add a post $p' \in \mathcal{P} \setminus \mathcal{X}$ that gives the greatest *marginal gain* until the budget is reached. This greedy step is formalized as,

$$p' = \operatorname*{arg\,max}_{p \in \mathcal{P} \setminus \mathcal{X}} \underbrace{\mathcal{R}(\mathcal{X} \cup \{p\}) - \mathcal{R}(\mathcal{X})}_{marginal\ gain\ \Delta_{\mathcal{X}}(p)}. \tag{6}$$

We term this straightforward greedy algorithm as PAS, short for *Post and Aspect Selection* algorithm. According to the theory developed by [Nemhauser *et al.*, 1978] on discrete optimization:

Lemma 1. *If $\mathcal{R}$ is a submodular, non-decreasing set function and $\mathcal{R}(\phi) = 0$, then a greedy algorithm finds a set $\mathcal{X}'$ which is no worse than constant fraction $(1 - 1/e)$ away from the optimal $\mathcal{X}^*$ where*

$$\mathcal{X}^* = \operatorname*{arg\,max}_{\mathcal{X} \in \mathcal{P} : |\mathcal{X}| \leq k} \mathcal{R}(\mathcal{X}).$$

It is clear that our $R(\mathcal{X})$ satisfies all the conditions of Lemma 1, meaning that the error rate of PAS algorithm is bounded by $1/e$. Although the algorithm works well in practice, it is rather inefficient. To show this, we now analyze its time complexity using big-O notations.

[1] When $\mathcal{X}$ is an empty set, we force $\mathcal{R}_i(\mathcal{X})$ to be zero to ensure the non-decreasing property.

Algorithm 1: FastPAS

```
   Input: Posts $\boldsymbol{P}$, Budget $\boldsymbol{k}$, Initial weights $\mathcal{W}$
   Output: Selected posts $\boldsymbol{X}$
1  $\boldsymbol{X} \leftarrow \emptyset$;   $\boldsymbol{U} \leftarrow \boldsymbol{P}$;
2  for $p \in \boldsymbol{U}$ do $\Delta_p \leftarrow \boldsymbol{R}(\{p\})$; $flag_p \leftarrow true$ ;
3  while $\boldsymbol{k} > 0$ and $\boldsymbol{U} \neq \emptyset$ do
       // Select one post
4      while true do
           // Get the top post
5          $p \leftarrow \arg\max_{p \in \boldsymbol{U}} \Delta_p$;
6          if $flag_p$ then
7              add $p$ to $\boldsymbol{X}$ and delete $p$ from $\boldsymbol{U}$;
8              update $w_i$ with decay if $a_i \in \mathcal{A}_p$;
9              break;
10         else
11             $\Delta_p \leftarrow \boldsymbol{R}(\boldsymbol{X} \cup \{p\}) - \boldsymbol{R}(\boldsymbol{X})$;
               $flag_p \leftarrow true$;
12         end
13     end
14     for $p \in \boldsymbol{U}$ do $flag_p \leftarrow false$ ;
15 end
```

Time Complexity Analysis. Let the number of total aspect candidates be M, and the number of averaged aspect candidates per post be $\overline{m}$. Suppose we have selected posts $\mathcal{X}$ and want to select the next post. Then the time cost for evaluating the marginal gain for one post p (*i.e.,* $\Delta_{\mathcal{X}}(p)$) is $O(\overline{m}M|\mathcal{X}|)$. Nevertheless, the key bottleneck of PAS is in the evaluation of Equation 6, which requires a traversal of all the unselected posts to find the best post in terms of $\Delta_{\mathcal{X}}(p)$. As such, for selecting k posts from the post collection $\mathcal{P}$, we roughly need to evaluate $\Delta_{\mathcal{X}}(p)$ for about $k|\mathcal{P}|$ times, which is unsuitable for large-scale data with over millions of posts.

4.2 FastPAS

The raw PAS algorithm treats each iteration as independent of each other, which makes it slow in selecting the post of largest marginal gain. Given the monotone property of our submodular objectives, we can achieve a significant speed-up by leveraging the dependency of two iterations to avoid unnecessary evaluation of posts' marginal gain. The idea is to re-use the marginal gain of previous iterations to assist in the post selection of the current iteration, rather than evaluating it for all unselected posts in each iteration [Wang *et al.*, 2017a]. Moreover, aspects captured in selected posts should be suppressed in future selection procedure. Therefore, assuming a_i has appeared in the selected posts set t times, its new weight will degrade to be $w_i e^{-t}$. Note that the monotone property of objectives still hold after degrading weights for some aspect terms. We give an example to illustrate how FastPAS works.

Assuming in the first iteration (the selected post set $\mathcal{X}_0$ is empty), we evaluated the marginal gain for all posts and obtained $\Delta_{\mathcal{X}_0}(p_2) > \Delta_{\mathcal{X}_0}(p_3) > \Delta_{\mathcal{X}_0}(p_1) > \Delta_{\mathcal{X}_0}(p_4)$. Clearly, p_2 would be selected in the first iteration, and the goal of the second iteration becomes selecting the post $p \in \{p_3, p_1, p_4\}$ that leads to the largest gain $\Delta_{\mathcal{X}_1}(p)$ where $\mathcal{X}_1 = \{p_2\}$. Instead of evaluating $\Delta_{\mathcal{X}_1}(p)$ for all the three remaining posts with updated $\mathcal{W}$, we would first check the value of p_3. If $\Delta_{\mathcal{X}_1}(p_3)$ is larger than $\Delta_{\mathcal{X}_0}(p_1)$, we would know that p_3 must be the one that has the largest gain, since the submodularity guarantees that $\Delta_{\mathcal{X}_0}(p_1) > \Delta_{\mathcal{X}_1}(p_1)$ (see Theorem 1). Thus the iteration can be stopped earlier without evaluating the marginal gain for p_1 and p_4.

The above example illustrates the key ingredient of our FastPAS for acceleration, which is detailed in Algorithm 1. The variable $\boldsymbol{X}$ stores the selected posts and $\boldsymbol{U}$ stores the remaining posts. For each post p, we store a marginal gain score Δ_p, which can be the latest score of the current status (labeled by $flag_p$ being true) or an old score of previous iterations (labeled by $flag_p$ being false). In each iteration, to select the post of largest marginal gain of the current status, we first check the post with the largest Δ_p (line 5): if Δ_p is the latest score (*i.e., $flag_p$* is true), we get the expected post and can perform early termination (line 7); otherwise, we need to update Δ_p to the latest score of the current status (line 9), and continue the process (line 5–10) until the expected post is found. Moreover, we use priority queue as the data structure for the variable Δ, which is rather fast in selecting the maximum element (*i.e.*, line 5).

Our FastPAS algorithm is guaranteed to return the same results as PAS, while being much faster than PAS owing to the early termination pruning. Analytically, adding a new post to $\mathcal{X}$ will not decrease $\Delta_{\mathcal{X}}(p)$ too much. As such, the top few posts of one iteration are very likely to be the top elements for the next iteration. Thus, FastPAS only needs to evaluate the marginal gain for these top few elements, rather than all the elements. In the later section, we empirically show that the acceleration is over $1,000$ times on average for our Twitter datasets consisting of $12,000$ tweets.

5 Experiments

As our proposed method aims to efficiently extract representative aspects and posts for brand monitoring, we list the research questions to guide the remainder of this section.

RQ1 Does our method extract salient aspects of the corpus? Whether monitoring competing brands is useful or not?

RQ2 How does our method perform as compared to other state-of-the-art methods?

RQ3 How is the efficiency of our accelerated FastPAS as compared to the raw PAS method?

5.1 Data and Experimental Settings

As there are no public benchmarks available for the brand monitoring task, we constructed the dataset by ourselves as in Table 1 . We issued queries "Singtel" and "StarHub" (which are the largest two telecommunication companies in Singapore) to Twitter search, and collected the tweets that were posted between September 2015 and March 2016. We carry out experiments on the two datasets separately. For aspect candidate extraction, we applied the Stanford parser to extract noun and noun phrases, followed by a modest filtering on the aspect candidates with threshold 10 (*i.e.*, appear in less than 10 posts) to combat the possible noises.

To quantitatively evaluate the extracted aspects, we adopt the metric *Normalized Discounted Cumulative Gain*

Table 1: Statistics of datasets for evaluation.

	Tweet#	Retweet#	Like#
Singtel	12,457	6,996	5,961
StarHub	12,726	23,308	14,024

(NDCG), a widely used measure in information retrieval community [He *et al.*, 2017]. It assigns higher importance to results at top ranks, scoring successively lower ranks with marginal fractional utility [Feng *et al.*, 2017]. For the relevance level, we invite five student volunteers to label each aspect with one of the three pre-defined levels: *Unrepresentative* (score 1), *Ordinary* (score 2) and *Representative* (score 3). Specifically, we ask the volunteers to rate the top-10 aspects produced by each method and report the averaged performance of NDCG@10 (the Fleiss' kappa value between them is 0.67). Since it is impractical to ask the volunteers to read through all tweets to judge the notion of representativeness, we apply a simplified process to reduce their workloads: we first provide them Annual Financial Report and Wikipedia page of the two companies, and then show them 5% of randomly sampled tweets.

5.2 Salient Aspects (RQ1)

Our method is designed to select both influential and representative posts and aspects. In Table 2, we show the top 5 posts selected by FastPAS for Singtel and highlight aspects.

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.*

As we can see, the selected posts show the important aspects for a telecommunication company and the detected aspects are rather diverse. The only redundant information comes from "singtel wifi", which occurs twice in the top five posts (both the 2^{nd} and 4^{th} post). Even so, we think this level of redundancy is generally acceptable since the 4^{th} post also contains another important aspect "mobile data" in addition to "singtel wifi". Moreover, if we increase the budget size to 10, the algorithm can additionally discover aspects such as "singtel's m-wallet", "cloud service", "net profit" and so on.

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*

We also show the top five posts for StarHub (Table 3), which is one of the main competitors of Singtel. We observe that the selected posts cover different aspects of StarHub, including data usage, sms, customer service *etc.* In contrast to the Singtel corpus, the top posts show that users complain a lot about data usage, which is evidenced by the negative sentiment of posts. Looking into the data, we find that most of the complaints are caused by one event — during the Chinese New Year, StarHub did not provide free data promotion but Singtel did. Interestingly, when we scrutinize the posts of Singtel, we only see very few users praise Singtel's data promotion. The reason might be that users are more likely to post complaints on social media when they have unsatisfactory experience. Thus, 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.

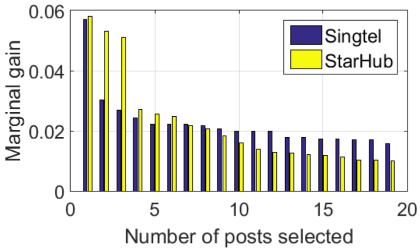


Figure 2: *Marginal gain w.r.t.* number of posts selected.

To verify that our definition of *representativeness* (*i.e.,* Equation 4) exhibits the submodular property, we show the *marginal gain* (short for MG) of each iteration (selecting one post) in Figure 2. First and foremost, we can see that MG decreases with more posts selected. For Singtel, the largest MG is obtained when selecting the first post, and the value drops dramatically in the subsequent selections; similar trend can be observed for StarHub. Since MG can be seen as the "gradient" of representativeness, the curve of MG indicates that the representativeness increases gradually with a decaying speed. This trend justifies the diminishing return property of our design.

5.3 Performance of Aspect Extraction (RQ2)

We now study the performance of aspect extraction by comparing with the following baselines:

AspectFreq Aspects are ranked by their frequency judged by number of occurrence. This baseline benchmarks performance by extracting the popular aspects of a corpus.

PostImpact This method first selects posts by their impact, and then identifies top aspects from the selected posts.

ASUM[Jo and Oh, 2011] It is a probabilistic topic model to jointly extract aspects and sentiments. We manually annotate the most representative word to indicate the aspect.

THUS[Long *et al.*, 2009] This method selects sentences (posts) that have least information distance to the corpus.

BMSF[Lin and Bilmes, 2010] This method summarizes the corpus by maximizing a non-monotone submodular function under a budget constraint.

Table 4 shows the performance in terms of $NDCG@10$ for aspect extraction. As can be seen, our method achieves the best performance on both datasets, outperforming other methods by a large margin. This verifies the rationality of our design, such that the extracted aspects are highly representative of the corpus.

Table 4: Aspect extraction evaluated by $NDCG$@10.

	Singtel	StarHub
AspectFreq	0.4003	0.4855
PostImpact	0.3523	0.4107
ASUM	0.5510	0.5526
THUS	0.5861	0.5367
BMSF	0.6464	0.5785
FastPAS	**0.8250**	**0.6757**

Among the baselines, ASUM is sophisticatedly designed for aspect extraction; however it does not take the impact of a post into consideration, and thus can be suboptimal in identifying representative aspects from social media posts. Similarly, both THUS and BMSF exhibit the same issue due to the overlook of social impacts — although they successfully extract aspects "3g" and "internet" as top aspects, they fail to identify the influential aspects like "free-data" that attract a large number of likes and retweets. PostImpact considers the impact information; however, its weak performance indicates that purely relying on the impact is insufficient. For example, the post promoting "2GB DATA Plan" attracts many retweets and likes in the Singtel dataset while this aspect lacks representativeness of the overall corpus.

Redundancy Analysis

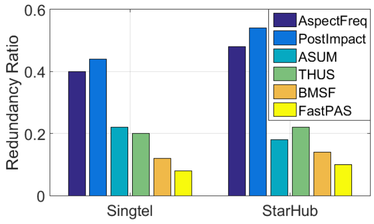


Figure 3: Redundancy ratio of the top 10 aspects. (RQ2)

In this evaluation, we invite the volunteers to label each of the top 10 aspects as redundant or not, and report the averaged redundancy ratio in Figure 3. It can be seen that our FastPAS method achieves the lowest redundancy ratio on both datasets — on average, only one of the top ten aspects is labeled as redundant by volunteers. We attribute the low redundancy to our two designs — the fuzzy matching scheme in measuring aspect similarity and the submodular property in selecting posts. Specifically, by using the fuzzy matching technique, our method manages to consider the similarity among aspects in the character-level, which helps to avoid redundancy like "singtel wi-fi" and "singtel's wifi". Moreover, by endowing the representativeness function with submodularity, our method can avoid selecting posts that contain similar aspects by suppressing the marginal gain. For those baselines, the relatively high redundancy ratio of AspectFreq and PostImpact (over 40%) indicates that neither the popularity of aspect nor the impact of post is sufficient to extract non-redundant aspects. ASUM achieves a lower redundancy ratio, despite that ASUM is not explicitly designed to reduce redundancy. This is because the clustering nature of topic models helps to discover aspects (topics) that are distinct from each other by modeling the co-occurrence statistics. However, it requires human efforts to manually label a topic with an aspect term. The summarization methods THUS and BMSF also achieve a relatively low redundancy, which is as expected since both methods are designed to reduce redundancy to a certain extent (THUS defines a conditional information distance for selecting posts, and BMSF uses a submodular function that is similar to MMR). The superior performance of our FastPAS over THUS and BMSF further justifies FastPAS's efficacy in extracting representative and non-redudant aspects.

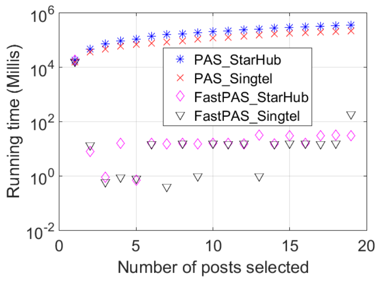


Figure 4: Running time of each iteration. (RQ3)

5.4 Speed Up (RQ3)

To show the speedup of our FastPAS method over the vanilla PAS, we compare the actual running time of the two algorithms. We implement the algorithms in Java and run them on the same machine (Intel i7 CPU of 3.60GHZ and 32GB RAM) in a single-thread for a fair comparison on efficiency.

Figure 4 shows the time consumption of each iteration (that selects one post) of the two methods on both datasets. The y-axis is in log-scale to better highlight the difference. As we can see from the figure, FastPAS significantly reduces the running time starting from the 2nd iteration. — on average, PAS requires over $10,000$ milliseconds to select one post, while FastPAS only takes less than 100 milliseconds. With more posts selected, both methods show an increasing trend of running time, which justifies our time complexity analysis.

6 Conclusion and Future Work

We addressed the problem of extracting representative aspects and posts from social media streams. We presented a new method — FastPAS — to meet specific requirements for brand monitoring. Extensive experiments showed that our method can extract representative aspects and posts efficiently, and meanwhile could control the redundancy well. In future, we will consider how to apply extracted aspects to downstream applications, such as using them to improve personalized summarization [Ren *et al.*, 2013] and the explainability of recommendation [Wang *et al.*, 2017b]. We will further improve the efficiency of our method to make it more suitable for practical use, for example by developing hashing techniques [Zhang *et al.*, 2016b]. Lastly, as social media is now being overwhelmed by multi-media content, such as image tweets and micro-videos [Chen *et al.*, 2017], it is interesting to unify such multi-modal data to perform more comprehensive aspect analysis.

Acknowledgments

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

References

[Carbonell and Goldstein, 1998] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, 1998.

[Chaudhuri *et al.*, 2003] Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, and Rajeev Motwani. Robust and efficient fuzzy match for online data cleaning. In *SIGMOD*, pages 313–324, 2003.

[Chen *et al.*, 2016] Tao Chen, Xiangnan He, and Min-Yen Kan. Context-aware image tweet modelling and recommendation. In *MM*, pages 1018–1027, 2016.

[Chen *et al.*, 2017] Jingyuan Chen, Hanwwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with feature- and item-level attention. In *SIGIR*, 2017.

[Feige *et al.*, 2011] Uriel Feige, Vahab S Mirrokni, and Jan Vondrak. Maximizing non-monotone submodular functions. *SJC*, pages 1133–1153, 2011.

[Feng *et al.*, 2017] Fuli Feng, Liqiang Nie, Xiang Wang, Richang Hong, and Chua Tat-Seng. Computational social indicators: a case study of chinese university ranking. In *SIGIR*, 2017.

[Glance *et al.*, 2005] Natalie Glance, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD*, pages 419–428, 2005.

[Haruechaiyasak *et al.*, 2013] Choochart Haruechaiyasak, Alisa Kongthon, Pornpimon Palingoon, and Kanokorn Trakultaweekoon. S-Sense: A sentiment analysis framework for social media sensing. In *IJCNLP*, page 6, 2013.

[He *et al.*, 2012] Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. Document summarization based on data reconstruction. In *AAAI*, pages 620–626, 2012.

[He *et al.*, 2015] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. TriRank: Review-aware explainable recommendation by modeling aspects. In *CIKM*, pages 1661–1670, 2015.

[He *et al.*, 2016] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR*, pages 549–558, 2016.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW*, pages 173–182, 2017.

[Hu and Liu, 2004] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[Jakob and Gurevych, 2010] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *EMNLP*, pages 1035–1045, 2010.

[Jo and Oh, 2011] Yohan Jo and Alice H Oh. Aspect and sentiment unification model for online review analysis. In *WSDM*, pages 815–824, 2011.

[Kobayashi *et al.*, 2007] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP*, pages 1065–1074, 2007.

[Liao *et al.*, 2014] Lizi Liao, Jing Jiang, Ying Ding, Heyan Huang, and Ee-Peng Lim. Lifetime lexical variation in social media. In *AAAI*, pages 1643–1649, 2014.

[Liao *et al.*, 2016] Lizi Liao, Qirong Ho, Jing Jiang, and Ee-Peng Lim. Slr: A scalable latent role model for attribute completion and tie prediction in social networks. In *ICDE*, pages 1062–1073, 2016.

[Lin and Bilmes, 2010] Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *NAACL*, pages 912–920, 2010.

[Liu *et al.*, 2015] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. Automated rule selection for aspect extraction in opinion mining. In *IJCAI*, pages 1291–1297, 2015.

[Long *et al.*, 2009] Chong Long, Minlie Huang, Xiaoyan Zhu, and Ming Li. Multi-document summarization by information distance. In *ICDM*, pages 866–871, 2009.

[Lovász, 1983] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. 1983.

[Moghaddam and Ester, 2012] Samaneh Moghaddam and Martin Ester. Aspect-based opinion mining from product reviews. In *SIGIR*, pages 1184–1184, 2012.

[Nemhauser *et al.*, 1978] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, pages 265–294, 1978.

[Paul and Girju, 2010] Michael Paul and Roxana Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*, pages 545–550, 2010.

[Popescu and Etzioni, 2005] Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In *EMNLP*, pages 339–346, 2005.

[Ren *et al.*, 2013] Zhaochun Ren, Shangsong Liang, Edgar Meij, and Maarten de Rijke. Personalized time-aware tweets summarization. In *SIGIR*, pages 513–522, 2013.

[Ren *et al.*, 2016] Zhaochun Ren, Oana Inel, Lora Aroyo, and Maarten de Rijke. Time-aware multi-viewpoint summarization of multilingual social text streams. In *CIKM*, pages 387–396, 2016.

[Wang *et al.*, 2015] Meng Wang, Xueliang Liu, and Xindong Wu. Visual classification by l1-hypergraph modeling. *IEEE TKDE*, pages 2564–2574, 2015.

[Wang *et al.*, 2017a] Meng Wang, Weijie Fu, Shijie Hao, Hengchang Liu, and Xindong Wu. Learning on big graph: Label inference and regularization with anchor hierarchy. *IEEE TKDE*, pages 1101–1114, 2017.

[Wang *et al.*, 2017b] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. Item silk road: Recommending items from information domains to social users. In *SIGIR*, 2017.

[Zhang *et al.*, 2016a] Hanwang Zhang, Xindi Shang, Huanbo Luan, Meng Wang, and Tat-Seng Chua. Learning from collective intelligence: Feature learning using social images and tags. *TOMM*, page 1, 2016.

[Zhang *et al.*, 2016b] Hanwang Zhang, Fumin Shen, Wei Liu, Xiangnan He, Huanbo Luan, and Tat-Seng Chua. Discrete collaborative filtering. In *SIGIR*, pages 325–334, 2016.