Discovering Topical Aspects in Microblogs

Abhimanyu Das
Microsoft Research
abhidas@microsoft.com

Anitha Kannan
Microsoft Research
ankannan@microsoft.com

Abstract

We address the problem of discovering topical phrases or "aspects" from microblogging sites like Twitter, that correspond to key talking points or buzz around a particular topic or entity of interest. Inferring such topical aspects enables various applications such as trend detection and opinion mining for business analytics. However, mining high-volume microblog streams for aspects poses unique challenges due to the inherent noise, redundancy and ambiguity in users' social posts. We address these challenges by using a probabilistic model that incorporates various global and local indicators such as "uniqueness", "diversity" and "burstiness" of phrases, to infer relevant aspects. Our model is learned using an EM algorithm that uses automatically generated noisy labels, without requiring manual effort or domain knowledge. We present results on three months of Twitter data across different types of entities to validate our approach.

1 Introduction

Microblogging sites such as Twitter and Weibo are evolving into the social platforms of choice for users to express and discuss, in real-time, their thoughts and ideas on a plethora of subject matters. It is thus important to use these microblog streams to identify the "buzz" or "talking points" regarding any topic or entity of interest, including organizations, products and social issues. This has several applications: For businesses, identifying what its customers are mostly talking about allows them to better engage with their customer base (Burton and Soboleva, 2011; Patino et al., 2012), fine-tune brand awareness and marketing campaigns (Popescu and Jain, 2011), and provide real-time feedback about customer preferences and complaints. Similarly, policy makers and think tanks would benefit from understanding the buzz around various socio-cultural or environmental issues, that could enable them to make well-informed choices and decisions. Infering such key talking points in social media also enables higher layer social-analytics applications such as trend detection, event tracking, and fine-grained opinion and sentiment analysis.

The goal of this paper is to automatically infer such entity-specific buzz in social media, which we represent using key phrases identified from microblog posts about the entity. Following past literature (Kobayashi et al., 2007; Mukherjee and Liu, 2012), we call these topical phrases as *aspects*. Thus, given a stream of microblog posts about an entity of interest, we devise an algorithm that *automatically discovers a ranked list of the top aspects that succinctly represent the buzz or key talking points among users about the entity.* As an illustrating example, Figure 1 shows the top 10 aspects discovered for each month by our aspect discovery algorithm for the Microsoft Surface tablet using 6 month of Twitter data. For each month we depict the key events and news stories (below the timeline) related to the Surface, along with the set of discovered aspects (above the timeline). As seen from the figure, several of the top aspects do not reflect product features or attributes, but instead capture the buzz among Twitter users around recent events or news related to the Surface. For example, the aspects "surface pricing" and "surface preorders" in October refer to the discussions on Twitter following a press release providing details of the Surface pricing and preorder dates. Similarly, the aspect "Oprah tweets" in November corresponds

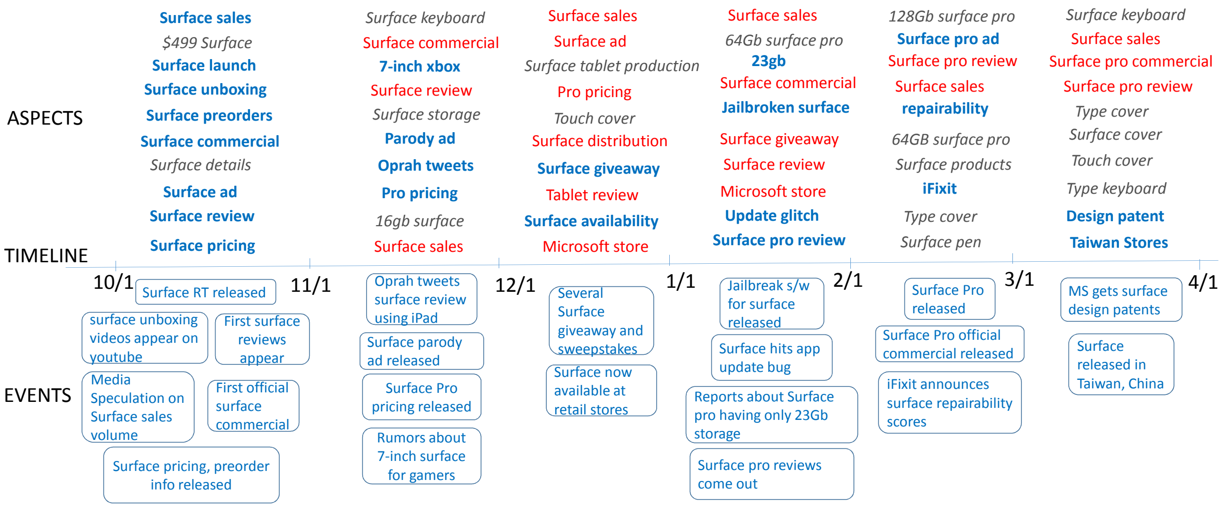


Figure 1: Temporal evolution of top monthly aspects for 'Microsoft Surface' over 6 months (October 2012 to March 2013). The aspects identified by our algorithm (for each month) is shown above the time line, while key events regarding the product is shown below. Aspects that directly relate to the events are shown in bold blue, while aspects that have no bearing to the news are shown in italics. The aspects that were related to an event in previous months but has persisted as an aspect are shown in red using normal font.

to discussions around media coverage of how Oprah Winfrey tweeted a review of the Surface tablet from her iPad. The aspects "iFixit" and "repairability" refer to the unveiling of the iFixit repairability report for the Surface in February. On the other hand, we also see more traditional product feature or attributes that are not correlated to external events but are key discussion points across multiple months, such as "keyboard" or "touch cover".

While there exists a rich line of work (refer to (Liu, 2012) for a comprehensive survey) in aspect identification from customer reviews, blogs or discussion forums, mostly for fine-grained opinion mining for products, there has been little work in the context of aspect discovery from large scale microblog posts. Microblogs pose a unique set of challenges that makes it difficult to directly apply existing methods from prior work. For example, in several papers, frequently occurring noun phrases is used as the building block for detecting aspects (Hu and Liu, 2004a; Hu and Liu, 2004b; Ku et al., 2006). However, for microblogs, frequency of a noun phrase alone is an insufficient indicator of an aspect, due to the inherent noise (unlike reviews, microblog posts are short and often not as focused) and redundancy (*e.g.,* due to retweeting in the context of Twitter). Yet another challenge unique to microblog streams is that the brevity of the posts provide inadequate context and structure. In addition, they are also noisy, with a single tweet often containing both relevant and irrelevant content for a given entity. Due to these reasons, well-known probabilistic approaches (*e.g.,* Topic Models (Mei et al., 2007; Titov and McDonald, 2008) or Conditional Random Fields (Jakob and Gurevych, 2010)), that work well for aspect identification from reasonably long and syntactically well-formed documents such as reviews and blogs, becomes immediately inappropriate in the microblog setting. Additionally, the high volume and velocity of social media streams calls for a scalable, fully automated approach that seamlessly works for a variety of entities and requires no domain-specific knowledge.

We address these challenges inherent in aspect discovery from microblog streams in a principled way: we propose quantifiable indicator measures of "uniqueness", "diversity" and "burstiness" based on insights that are fairly intuitive and yet are generic enough to model the characteristics of relevant aspects for a range of diverse entities. We represent candidate phrases in terms of these three indicators. We propose a probabilistic model for scoring the candidate phrases (§ 2.3) corresponding to an entity of interest. For every entity, the model automatically clusters the indicators and for each cluster, learns relative importance between the indicators for scoring the candidate aspects. Given a collection of <candidate aspects, noisy label> pairs where the noisy label reflects if the corresponding candidate is an aspect (albeit, noisly), we use an Expectation-Maximization algorithm for training the model. We also present an approach to leverage web search engine results (§ 2.4) to automatically obtain noisy labelled data for any entity. While being entity specific, our approach is highly scalable, entity-agnostic and does not require any manual effort. We validate our results on diverse entities, using *all* tweets from Twitter corresponding to a three month period from January 2013 to March 2013.

Related Work: To the best of our knowledge, the only works related to aspect discovery from microblog posts are (Spina et al., 2012) and (Zhao et al., 2011). In (Spina et al., 2012), four information retrieval functions were compared for identifying aspects from a set of tweets about companies. They showed that a TF-IDF based approach performed the best. Their experiments were however not performed across multiple domains, and used a very small number of tweets for each company. Furthermore, our 'uniqueness' based ranking (§ 2.2) that we use as one baseline is quite similar to their TF-IDF approach, and in our large scale evaluation over diverse domains, we show that TF-IDF or uniqueness alone is not sufficient for efficient aspect discovery (§ 3). The work by (Zhao et al., 2011) proposes an unsupervised approach for keyphrase ranking based on measures of "interestingness" (which is similar to our uniqueness indicator) and "relevance". However, as we show in our experiments (§ 3), the performance of this method is entity–dependent and does not naturally scale to all entities.

The rest of the paper is organized as follows. We describe our algorithm in § 2, including the various indicators that we use to characterize an aspect, the automatic label generation, and our probabilistic model. In § 3, we present experimental results and evaluation of our algorithm along with other baselines on the three month Twitter data set. We conclude in § 4 with remarks on future work.

2 Approach

We formulate the problem of identifying aspects as follows: **Problem statement:** Let e be an entity and s be a time period of interest. We use $\mathcal{T}^s$ to denote the set of all tweets in time period s, and $\mathcal{T}_e^s \subset \mathcal{T}^s$ to be the set of all tweets about e in time period s [1]. Then, we wish to identify the set of k phrases from $\mathcal{T}_e^s$ which are most likely to be valid aspects of e.

Solution overview: Given e, we first identify a set of candidate phrases for aspects from $\mathcal{T}_e^s$ (§2.1). For each phrase, we compute a global indicator, *uniqueness*, that measures how strongly the phrase is correlated with e by comparing its occurrence in $\mathcal{T}_e^s$ and $\mathcal{T}^s$. We also compute two local indicators, *diversity* that measures how diversely the phrase is used in $\mathcal{T}_e^s$, and *burstiness* that measures the temporal activity around the phrase usage in $\mathcal{T}_e^s$ (§ 2.2). We train a probability model (§ 2.3) that captures non-linear relationships between the indicators using a combination of linear decision surfaces. The training labels are obtained using a completely automated approach (§ 2.4). The model is trained independently for each entity, and subsequently used in inferring aspects for the entity during the time period of interest.

2.1 Candidate Aspects

We expect an aspect of an entity to be a phrase on which users can say something subjective. This intuitive requirement is enforced by restricting candidate aspects to be noun phrases (Hu and Liu, 2004b; Popescu and Etzioni, 2007) that are qualified with an adjective within short proximity (around four words) in at least one tweet (Blair-Goldensohn et al., 2008). We use a Twitter-specific part-of-speech tagger (Owoputi et al., 2013) to identify a candidate set of noun phrases in $\mathcal{T}_e^s$ that are used in conjunction with an adjective. After resolving plural nouns to their singular forms, this results in a few thousand candidate phrases per entity for a month of tweets.

2.2 Indicators

We represent a phrase using measurements across three dimensions that captures "diversity", "uniqueness" "burstiness" of usage. These are described in detail below.

[1] While accurately classifying microblog posts to extract posts relevant to an entity is a research problem in itself, this is outside the scope of this work. In this work, we use keyword based classifiers for our entities.

2.2.1 Diversity

Intuitively, a genuine aspect of an entity is more likely to have been discussed on Twitter in the context of that entity, compared to other noun phrases. While one can consider a metric like occurrence frequency (*e.g.,* (Liu, 2012)) to capture this intuition, in microblog settings like Twitter, this can overestimate the importance of a phrase because of redundancy of content due to (a) simple retweeting by followers, (b) multiple users posting the same or very similar content, especially when talking about news and events, and (c) same user posting multiple versions of the same tweets due to automated tweet applications. As an example, the most frequently used noun phrase on Twitter for the entity 'Microsoft Surface' during March was "tablet-a-day giveaway". However, all the tweets containing this phrase referred to a lottery contest that required users to tweet a pre-specified sentence about the Surface. Hence, this phrase cannot be considered a relevant aspect.

We propose factoring out such redundancy by using a notion of "diversity" of content about that aspect. To this effect, for each candidate aspect, its "Diversity" indicator is obtained by computing a score based on the amount of diverse content in the set of tweets about the aspect. To efficiently compute this diversity score, we use the Simhash algorithm (Charikar, 2002) based on Locality Sensitive Hashing (Indyk and Motwani, 1998). Simhash measures the similarity of two tweets t_1 and t_2 by hashing them into small f-bit fingerprints (we use $f = 128$), and comparing the Hamming distance between them. The Locality Sensitive Hash function H used by Simhash ensures that

$$Pr[H(t_1) = H(t_2)] = Sim(t_1, t_2),$$

where $Sim(t_1, t_2)$ is the cosine similarity between t_1 and t_2.

Thus, it suffices to compute a diversity score on the (much smaller) set of 128-bit fingerprints of the tweets containing the aspect. We define this score as the cardinality of the largest subset $S \subset \mathcal{T}_e^s$ of tweets such that the Hamming distance d between the fingerprints of any pair in S is at most 90% of the fingerprint length. While this is a combinatorially hard problem, we use a greedy heuristic to approximate this score using the following steps: 1) Initialize S to a random tweet $r \in \mathcal{T}_e^s$. 2) At each iteration, let $t \in \mathcal{T}_e^s \setminus S$ maximize $D(S, t)$. (Here we define $D(S, t) = \min_{x \in S} d(H(x), H(t))$). If $D(S, t) > 0.9$, add t to S. Else return $|S|$.

2.2.2 Uniqueness

Another property of a relevant aspect for an entity is a notion of "uniqueness" to that entity. Intuitively, an aspect should have a higher propensity of being used in tweets about that entity, compared to a generic set of tweets. For example, in the case of Microsoft Surface, several commonly used noun-phrases might have a high frequency of occurrence or a high diversity score such as "news" or "store". However such phrases are arguably too generic to be considered as an aspect of Microsoft Surface. Hence we need to evaluate a candidate noun phrase in terms of its frequency in the set of tweets for that entity, versus its frequency across all tweets in the same time window. In particular, we define the uniqueness indicator of a phrase p in a time period s as:

$$\text{uniqueness}_e^s[p] = \frac{\sum_{t \in \mathcal{T}_e^s} I[\mathrm{p} \in t]}{\sum_{t \in \mathcal{T}^s} I[\mathrm{p} \in t] + \theta}, \tag{1}$$

where $I[\mathrm{p} \in t]$ is an indicator that evaluates to 1 if the tweet t contains the phrase p, and 0 otherwise. θ enforces minimal support (θ tweets from $\mathcal{T}_e^s$) required for p to be unique. We used $\theta = 10$.

Note that this is reminiscent of the tf-idf metric in information retrieval and also used in (Spina et al., 2012); The numerator corresponds to the notion of term-frequency and the denominator to document frequency. This can also be interpreted probabilistically, by considering a bernoulli variable Z that models how unique the phrase is to e. Then the above definition is similar to a maximum-likelihood estimate of Z using a Beta distribution with θ as the prior.

2.2.3 Burstiness

Another indicator of a relevant aspect of an entity is a noun phrase that has an unexpected surge in its frequency of occurrence among tweets of the entity, in a short period of time. This could be due to an emerging news story, event or talking point about the entity and hence indicate that the phrase is strongly related to the entity, even if the overall frequency of the phrase over a larger time period might be low.

We capture this notion using the "burstiness" indicator. For each candidate noun-phrase, we create a time-series of its occurrences in tweets of the entity within the specific time window. We then use the burst model due to Kleinberg (Kleinberg, 2002) to extract a burstiness score for the noun-phrase. Kleinberg's model uses a finite-state automaton with different states corresponding to different emission frequencies, where state transitions from a low-frequency state to a high-frequency state signify the onset of a burst. We use an R-implementation (url, 2014) of this algorithm on the time series of occurrences of a noun-phrase to identify the corresponding burst levels, and define the burstiness score of the noun-phrase as the sum of these burst levels. For example, the aspect "shipping lanes" detected by our algorithm for the entity "Global Warming" has relatively low frequency of occurrence overall, however it was a topic of intense discussion on Twitter during a week when mainstream news media reported on a PNAS article discussing opening up of new shipping lanes through the Arctic ocean due to global warming(url, 2013).

2.3 Probabilistic model for aspect identification

Given a candidate phrase and its measurements of indicators, we would like to rank these based on a learned model that takes into account these varied interactions between the indicators. One approach is to directly train a linear classifier such as logistic regression using a training set of <phrase,binary label> pairs. As we show in § 3, this approach does not capture the non-linear dependencies among the indicators and the label, resulting in poor performance. In this paper, we jointly model the space of indicator variables and their labels, which we describe next.

2.3.1 Model specification

A candidate phrase is represented by a three-dimensional continuous-valued random variable $\mathbf{x}$, where x_1 corresponds to 'Diversity', x_2 to 'Uniqueness' and x_3 to'Burstiness'. The relationship between these indicators is captured by a probabilistic Gaussian mixture model. Let c be a random variable with discrete distribution over m components. Then,

$$p(\mathbf{x}, c) = p(c)p(\mathbf{x}|c) = \pi_c \mathcal{N}(\mathbf{x}|\mu_c, \Phi_c), \tag{2}$$

where $p(c)$ is a Multinomial distribution with probability π_c for the c^{th} component such that $\sum_c \pi_c = 1$ and $p(\mathbf{x}|c)$ is a Gaussian distribution for c^{th} component with mean vector μ_c and covariance matrix, Φ_c. Let y be the Bernoulli random variable representing whether a phrase is an aspect, such that:

$$p(y = 1|\mathbf{x}, c) = \frac{1}{1 + \exp(-(\mathbf{w}_c^T \mathbf{x} + b_c))}. \tag{3}$$

Then, the joint distribution over the variables is

$$p(\mathbf{x}, y, c) = p(c)p(\mathbf{x}|c)p(y|\mathbf{x}, c) \tag{4}$$

Note that unlike a mixture of logistic regressors (Bishop, 2007), this formulation captures $p(\mathbf{x}|c)$ which is central to modeling the correlation between the indicators. One can view our formulation as a variant of mixture of experts (Jacobs et al., 1991) wherein the gating functions are represented using the posterior over the mixture model components, as opposed to the soft-max function typically used.

2.3.2 Learning

Given a set of training examples,$[\mathbf{X}, \mathbf{y}] = \{\mathbf{x}_n, y_n\}_{n=1}^N$, the model parameters $\{\mu_c, \Phi_c, \pi_c, \mathbf{w}_c, b_c\}_{c=1}^K$ are learned so as to maximize the probability of observations, $p(\mathbf{X}, \mathbf{y})$. Assuming the training examples are independent and identically distributed, we use an Expectation-Maximization algorithm to learn parameters that maximize the probability of observations, $p(\mathbf{X}, \mathbf{y}) = \prod_{n=1}^N p(\mathbf{x}_n, y_n)$ or equivalently, its log:

$$\log p(\mathbf{X}, \mathbf{y}) = \sum_n \sum_c p(c|\mathbf{x}_n, y_n) \log \frac{p(c_n, \mathbf{x}_n, y_n)}{p(c|\mathbf{x}_n, y_n)} \tag{5}$$

$$= \sum_n \sum_c p(c|\mathbf{x}_n, y_n) \log \frac{p(c)p(\mathbf{x}_n|c)p(y_n|\mathbf{x}_n, c)}{p(c|\mathbf{x}_n, y_n)}, \tag{6}$$

where $p(c|\mathbf{x}_n, y_n)$ is the posterior distribution over c. The parameters are learned using the EM algorithm. by iterating between Expectation(E)-step in which $p(c|\mathbf{x}_n, y_n)$ is estimated for each training instance, and the Maximization(M)-step in which parameters of the model are estimated:

E-step: In this step, $p(c|\mathbf{x}_n, y_n)$ is computed for each training instance by taking derivative of eq. 6 with respect to $p(c|\mathbf{x}_n, y_n)$ and setting to zero, so that $p(c = j|\mathbf{x}_n, y_n) \propto p(\mathbf{x}_n, y_n, c = j)$.

M-step: The mixture component parameters (means and covariances) are updated as weighted averages and deviations from the mean, weighted by the posterior computed in the E-step:

$$\mu_c = \frac{\sum_{n=1}^{N} p(c = j|\mathbf{x}_n, y_n)\mathbf{x}_n}{\sum_{n=1}^{N} p(c = j|\mathbf{x}_n, y_n)} \qquad \Phi_c = \frac{\sum_{n=1}^{N} p(c = j|\mathbf{x}_n, y_n)(\mathbf{x}_n - \mu_c)(\mathbf{x}_n - \mu_c)^T}{\sum_{n=1}^{N} p(c = j|\mathbf{x}_n, y_n)} \tag{7}$$

The weight vector,$\mathbf{w}_c$, for each of the logistic component is estimated using the re-weighted least squares (IRLS) algorithm (Bishop, 2007):

$$\mathbf{w}_c = \arg\max_{\mathbf{w}} \sum_{n} p(c|\mathbf{x}_n, y_n) \log p(y_n|\mathbf{x}_n, c). \tag{8}$$

Scoring Function: Once the model is trained, it is used to score a candidate phrase for being an aspect. For any phrase with indicator vector $\mathbf{x}$, the probability of it being an aspect is given by the convex combination (weighted by $p(c|\mathbf{x})$) of the outputs from all the regressors: $p(y|\mathbf{x}) = \sum_c p(c|\mathbf{x})p(y|\mathbf{x}, c)$

Choice of number of components: The number of components m is a free parameter in our model, and its value is a function of the training dataset. We use Bayesian information criteria (BIC) (Schwarz, 1978) to choose the optimal number of components for training. In particular, we train models by varying K and pick the one with the largest BIC given by $\log p(\mathbf{X}, \mathbf{y}|\theta_m) - \frac{|\theta_m|}{2} \log N$ where θ_m is the model with m components having $|\theta_m|$ parameters, N is the number of data points and $\log p(\mathbf{X}, \mathbf{y}|\theta_m)$ for fixed θ_m is given by eq. 6.

2.4 Automatic generation of training data

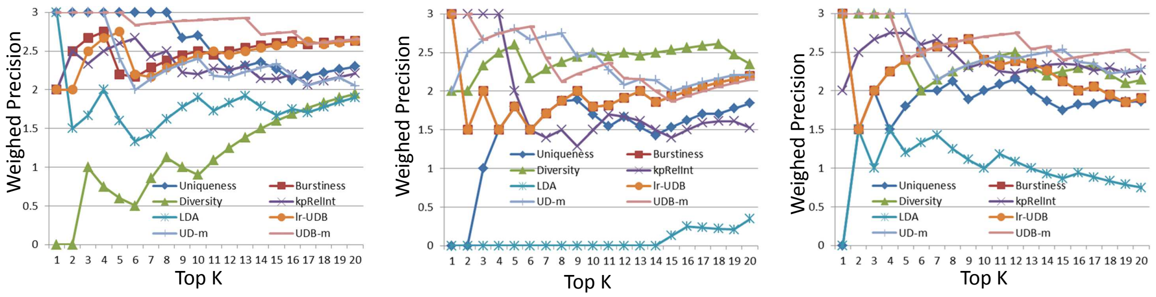
We use a fully automated approach to (noisily) label candidate phrases. The approach is based on the premise that a phrase that is related to the entity and is also popular on the web is more likely to be a potential aspect for the entity. We operationalize this by issuing each phrase to be labeled as a query to a web search engine and retrieve top 50 results. Then, we label it as an aspect if, at least 10% of the top 50 web results have web page titles that are relevant to the entity (determined by the same rules that is used for tweet classification (§ 3.1)) and all the unigrams in the phrase is contained in them.

This approach can result in noisy labels since a candidate phrase that have huge web presence may not be an aspect, and vice versa. In spite of this, we observed reasonable correlation between the propensity of a phrase on Twitter to be a true aspect and the quality of web search result that we can retrieve. Thus, this approach results in generating large noisily labeled datasets, which can often be more effective than a small dataset with high quality labels (Fuxman et al., 2009).

3 Evaluation

We compare our algorithm described in § 2.3 (which we denote *UDB-m*) against the following algorithms: (1) *kpRelInt*: the keyphrase ranking algorithm of (Zhao et al., 2011) applied to our candidate aspects, (2) *lr-UDB*: ranking based on probabilities obtained using a trained, single-component logistic regression model using uniqueness, diversity and burstiness indicators as features, (3) *UD-m*: ranking based on our probabilistic mixture model of § 2.3 where we only used Uniqueness and Diversity indicators but not Burstiness and (4) *LDA*: an approach based on training a 50 component Latent Dirichlet Allocation (Blei et al., 2003) on tweets of that entity, from which we then manually constructed aspects from the best 20 topics.

We also consider rankings based solely on the various indicator scores themselves: uniqueness (*U*), diversity (*D*) and burstiness (*B*), to understand the effectiveness of each of these indicators. Note that *U* is a stronger baseline than the TF-IDF metric (Spina et al., 2012).



(a) Windows 8

(b) Global Warming

(c) Microsoft Surface

Figure 2: Weighted Precision across entities

3.1 Dataset

We studied six entities from varied domains including products, environmental issues and personalities:“Windows 8”, “Microsoft Surface”, “Hyundai”, “Organic Food”, “Global Warming”, and “Tiger Woods”. We obtained the set of all English language tweets posted in a three month time period from Jan 1, 2013 to March 31, 2013. For each entity, we classified the tweets pertaining to that entity by using simple keyword-based classifiers. For instance, for the entity ‘windows 8’, the keywords corresponded to the set {”windows 8”, ”win8”, ”windows8”, ”win 8”, ”#win8”, ”#windows8”, ”#microsoftwindows8”}. In total, we obtained about three million English tweets across all the six entities that we used, with around $100,000$ to $800,000$ tweets for each entity.

Train/Test split: We used data from January to train *UDB-m*, *UD-m* and *lr-UDB* algorithms. We evaluated all algorithms on data from February and March, and obtained qualitatively similar results for both months. We present results only from March, due to space constraints.

3.2 Precision analysis of inferred aspects

The goal of this experiment is to obtain a precision measure for the various algorithms in inferring relevant aspects. Since it is impractical to manually create a ground truth test set of aspects for each entity by inspecting all the tweets, we take an approach used in (Spina et al., 2012). For each entity and month, we pooled together the top 20 aspects identified by all the algorithms under consideration. We used three judges in our organization as human assessors who manually annotated these candidate aspects on a 4-point relevance scale (with ‘3’ being most relevant and ‘0’ being irrelevant to the entity).

Metrics: Let S be the list of top K phrases identified as aspects by an algorithm, with $S[i]$ being the i^{th} phrase. For every phrase $p \in S$, let $R(p) \in [0, 3]$ be the average of the relevance rating provided by the three judges. Then, $\mathrm{WeightedPrecision}$ $@\mathrm{K}$ of the algorithm at the top K rank is given by $\frac{\sum_{i=1}^{K} R(S[i])}{K}$ (Sakai, 2007). Note that *Weighted Precision @ K* lies in the range [0,3] with higher values indicating that the list of top K aspects is more precise.

Results: Figure 3 shows the Weighted Precision at top K ranks ($K = 1, \ldots, 20$) for each algorithm, averaged across all the entities. For each algorithm and value of K, the marker size of each point in the plot is proportional to the variance in the algorithm’s weighted precision. Observe that *UDB-m* consistently has high Weighted Precision scores across all values of K and has the lowest variance, showing its efficacy in discovering aspects with high precision across all entities. Contrast this with the relatively poorer performance of *lr-UDB* that uses a simpler logistic regression model on the same three indicators. This highlights the importance of using a multiple-component mixture model (as opposed to a single component) to capture the non-linear dependencies among the three indicators for an entity. We discuss this further in § 3.4.

The next closest contender after *UDB-m* is *UD-m* that uses only the uniqueness and diversity indicators. The non-trivial gap between *UDB-m* and *UD-m* indicates the importance of incorporting burstiness. The *kpRelInt* algorithm of Zhao et al. (Zhao et al., 2011) actually performs quite poorly. We observed two reasons for this: first, the interestingness score in *kpRelInt* that is based on the ratio of retweets to tweets does not capture key aspects that may have been frequently used by tweets (but not necessarily retweeted often), and secondly it gives undue importance to words in tweets that are meant to be retweeted by design (for example, as part of a contest, announcement or giveaway). Indeed, the former reason is precisely addressed by our diversity indicator, whose importance is be seen from the fact that among all the three indicators, D performs the best.

We note that methods that use only one of the indicators (U, D and B) have large variance in their performance across entities emphasizing the entity-specific nature of these algorithms (we comment on this shortly) making them ill-suited for large scale domain-agnostic applications. Finally, we see that *LDA* performs the worst among our baselines, due to the inherent brevity, ambiguity and noise in tweets.

Entity-specific analysis: Consider Figure 2 that compares entity-specific performances of the algorithms considered. Figure 2a shows their performance for 'Windows 8'. For this entity, U, *UD-m* and *UDB-m* all perform equally well for small values of K, however the performance of *UDB-m* stays stable even for large values of K, while that of U and *UD-m* deteriorate. Contrast the relatively poor performance of B for Windows 8 with its performance for 'Global Warming' in Figure 2b. We see that the precision of *UDB-m*, which is still higher than most of the other algorithms, aligns with that of B for small K. This is due to the inherent nature of this entity, for which much of the chatter on Twitter tends to revolve around major news events. We discuss this in more detail in § 3.5. *UD-m*, which performed very well for the Windows 8 entity, does not have as good precision in this case, because it does not factor in this important effect of burstiness. Figure 2c shows the performance of the algorithms for the 'Microsoft surface' tablet. Again, *UDB-m* mostly outperforms the other methods across the range of K values, but is matched by *UD-m* and, to a lesser extent, by D for small K. Burstiness no longer plays such an important role - the tweets for Surface in March tend to be mostly comments on the features, commercials and accessories related to the product, and not so much related to news.

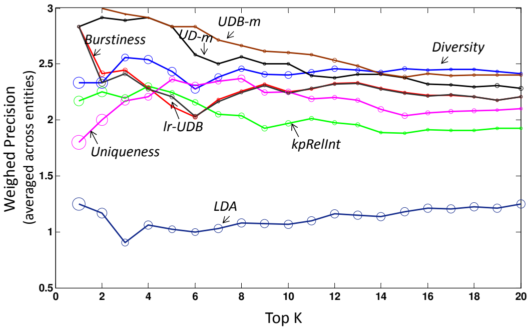


Figure 3: Average and variance (over all entities) for Weighted Precision for various algorithms. (Best viewed in color)

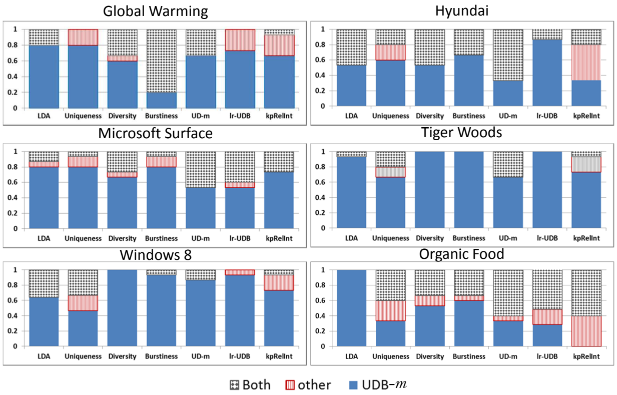
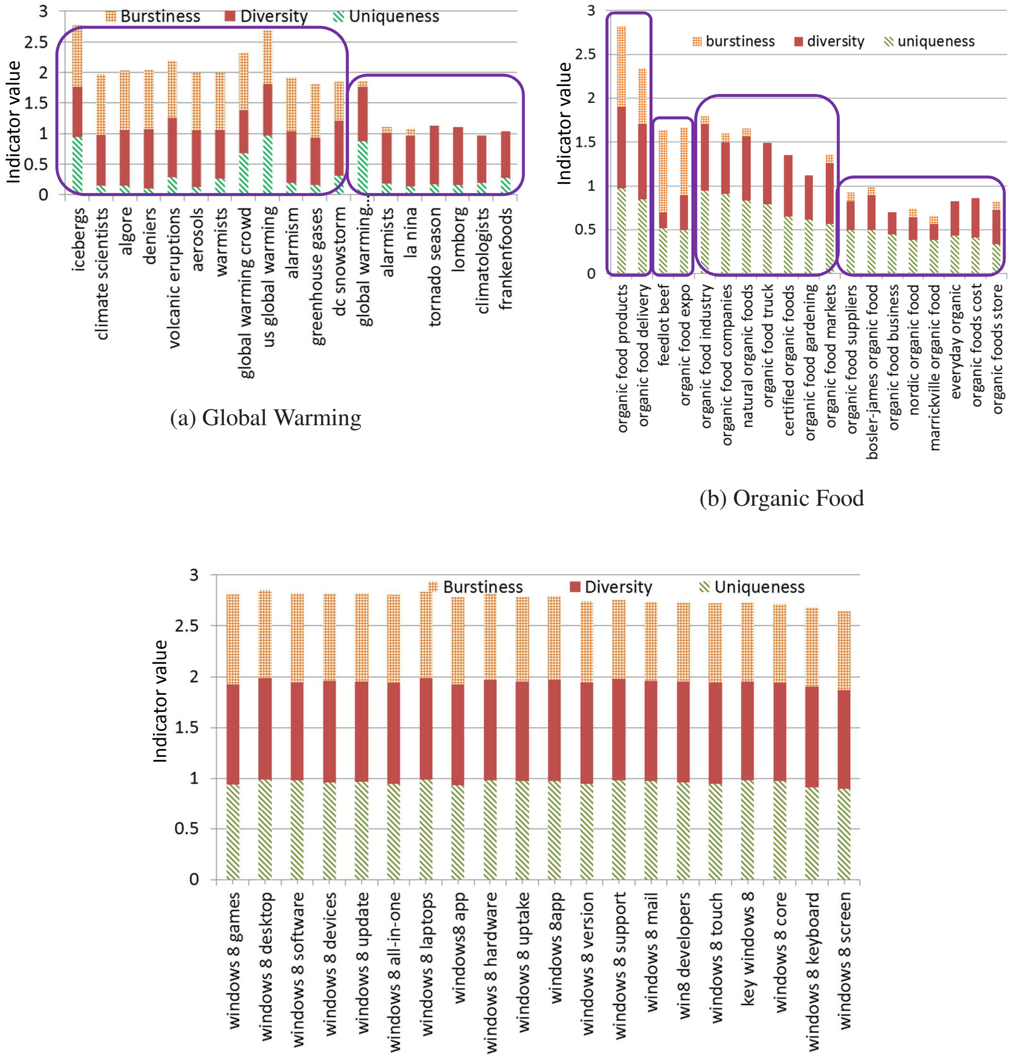


Figure 4: Pairwise preferences (at top 10) for *UDB-m* and other algorithms studied

3.3 Pairwise Preference comparison of inferred aspects

Here, we quantify the overall precision of the ranked list of aspects identified by various algorithms. We conduct pairwise evaluation using Amazon Mechanical Turk. Each Human Intelligence Tasks (HIT) consists of a pair of top 20 ranked aspect lists for an entity, with one list from *UDB-m* and the other chosen from one of the baseline algorithms. For each pair, we randomly permuted the order for each HIT (considered 5 random orderings). Each pair was judged by 5 judges, resulting in 25 judgments for each <entity, *UDB-m*, baseline-algorithm> triplet. Each judge was asked to study the two lists and specify which of the two was more relevant (or choose "Both are comparable"). Since the judges do not have access to the tweets, they were given instructions to perform a web search using the aspect and the entity name as a query string, restricted to the appropriate month. They were then asked to use the search results to guide them in determining which of the ranked lists was more relevant. We computed the Fleiss-κ inter-annotator agreement across each entity and method to be 0.68 on average, showing substantial agreement among the judges.



(a) Global Warming

(b) Organic Food

(c) Windows 8

Figure 5: Strengths of indicators for top 20 aspects (sorted according to the components)

Results: Figure 4 plots the results of the pairwise preference evaluations for all the entities for the month of March. Each of the seven bars corresponds to the preference results obtained by comparing *UDB-m* versus one of *lr-UDB*, *kpRelInt*, *LDA*, *UD-m*, *U*, *D* and *B*. For each pair, we plot the fraction of the 25 judges that 1) voted for *UDB-m* 2) voted for the other algorithm, or 3) answered “both are comparable”. We observe that for almost all the entities, the fraction of votes for *UDB-m* was higher than the votes for the other algorithms. The only exception was organic food for which *kpRelInt* performed better than *UDB-m*. There were some cases for which a majority of judges said both aspect lists were comparable. These cases mostly involved comparisons of *UDB-m* versus *UD-m* which suggests that after *UDB-m* the next best performing algorithm was *UD-m* (We had observed the same effect in precision analysis § 3.2). Another case where both aspect lists are comparable was *UDB-m* vs *B* for the entity ‘global warming’. As described in § 3.5, the buzz around this entity is often centered around news events and hence the top 10 aspects identified by our algorithm coincides closely with burstiness of the phrases. Hence, *UDB-m* and *B* perform quite comparably.

3.4 Importance of Multiple Components

UDB-m uses multiple components to model the data, with actual number of components inferred using BIC. Here, we demonstrate the importance of using varied number of components for each entity. Figure 5 shows the top 20 aspects identified for three of the entities. For ease of exposition, we group

Global Warming	**Hyundai**	**Microsoft Surface**	**Tiger Woods**	**Windows 8**	**Organic Food**
volcanic eruptions	hyundai santafe	surface keyboard	tiger woods commercial	windows 8 games	organic food delivery
deniers	hyundai sonata	surface sales	us skier	windows 8 desktop	organic food truck
aerosols	starex	surface pro commercial	tiger woods #2	windows 8 software	certified organic foods
tornado season	fuel cell	surface pro review	cadillac championship	windows 8 devices	organic food business
al gore	hyundai genesis	type cover	arnold palmer invitational	update windows 8	everyday organic
lomborg	tucson	surface review	tiger woods #3	windows 8 all-in-one	organic food gardening
global warming awareness	r-spec	touch cover	tiger woods number	windows 8 laptops	organic foods cost
us global warming	i-deal	type keyboard	tiger woods video	windows8 app	cafe bahrain
dc snowstorm	elantra	benchmarks surface	tiger woods house	windows 8 hardware	nordic organic food
icebergs	entourage	surface tablet line	skier lindsey vonn	windows 8 uptake	organic food industry

Table 1: Top 10 aspects identified by our algorithm for various entities

aspects list based on which component they belonged to (the one with largest posterior probability). For each aspect, we show a stacked bar representing the values for the three indicators (hence the maximum length of the bar is 3).

Consider Figure 5a corresponding to 'Global Warming'. Here, only a two component model was trained: one to model large values for diversity and burstiness, and another to model large values for diversity. While highly bursty and diverse aspects such as 'volcanic eruptions' are explained by the component that captures large diversity and burstiness values, aspects such as "tornado season" and "lomborg" that are widely discussed in diverse contexts but are not bursty are captured in another component.

Contrast this with Figure 5b for 'organic food'. This entity was automatically trained using a six component model out of which four participated in identifying the top 20 aspects. The aspect 'organic food products' from the first component has large values for all three indicators. In contrast, the aspects 'organic food truck' and 'organic food gardening' from the third component have high uniqueness and diversity values but low values for burstiness, indicating that they are consistently talked about through the month. The aspect 'feedlot beef' in the second component has low values for diversity, but has large values for burstiness indicating a spike in chatter around feedlot beef in the context of organic food.

Figure 5c shows the corresponding plot for 'Windows 8'. All the top aspects come from a single component. While one may be tempted to use only one component for this entity, our model used six components in order to explain the high variance in the training data. The remaining components were useful in weeding out noise. This can also be seen from the improved performance of *UDB-m* in comparison to *lr-UDB* which uses only a linear classifier (Figure 2a).

3.5 Qualitative Results

Table 1 shows the top 10 aspects identified by our algorithm for the month of March 2013. Consider the entity, 'global warming'. In Twitter, we found that discussions around this entity were highly news (or event) driven and this is reflected in the identified aspects. For instance, 'volcanic eruptions' and 'aerosols' corresponds to news reports of a study that showed how aerosols from modest volcanic eruptions may mask global warming effects. The aspect 'lomborg', referring to Bjorn Lomborg was the subject of much discussion in March; With his article in WSJ on heavy carbon–di–oxide emissions from electric cars charging, Lomborg created a stir among environmentalists.

In contrast, the top ranking aspects for Hyundai on Twitter corresponded mostly to chatter about various car models. Hyundai's announcement in March of their intention to offer fuel-cell cars in the US led to a lot of buzz around this topic, as aptly identified by the aspect 'fuel cells'.

There were three major events in March about Tiger Woods that created buzz on Twitter (and mainstream news media): his Cadillac Championship performance that led to his regaining the number one spot in golf, his relationship with the US skier Lindsey Vonn, and his rivalry with Graeme McDowell during the Cadillac championship. All these events are identified as aspects for the entity 'Tiger Woods'.

4 Concluding Remarks

In this paper, we studied the problem of inferring the key talking points or *aspects* about entities from microblog streams. We presented a probabilistic model to automatically infer these aspects from microblog streams for any specified domain, with *no* manual effort or domain knowledge about the entity.

We presented indicators such as “uniqueness”, “diversity” and “burstiness” to capture characteristics of aspects in the microblog context. Our large scale empirical evaluation over three months of Twitter data for entities from various categories validated the efficacy of our approach.

A key direction for future work is the problem of clustering semantically similar aspects pertaining to an entity (*e.g.,* ‘volcanic eruptions’ and ‘aerosols’ for ‘global warming’) to get a more succinct representation of the aspects. Another line of work is to leverage the temporality of these aspects in building temporal aspect discovery models.

References

[Bishop2007] Christopher Bishop. 2007. *Pattern Recognition and Machine Learning*. Springer.

[Blair-Goldensohn et al.2008] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Burton and Soboleva2011] Suzan Burton and Alena Soboleva. 2011. Interactive or reactive? marketing with twitter. *Journal of Consumer Marketing*, 28(7):491–499.

[Charikar2002] Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM.

[Fuxman et al.2009] Ariel Fuxman, Anitha Kannan, Andrew B Goldberg, Rakesh Agrawal, Panayiotis Tsaparas, and John Shafer. 2009. Improving classification accuracy using automatically extracted training data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1145–1154. ACM.

[Hu and Liu2004a] Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Hu and Liu2004b] Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

[Indyk and Motwani1998] Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM.

[Jacobs et al.1991] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

[Jakob and Gurevych2010] Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045. Association for Computational Linguistics.

[Kleinberg2002] Jon Kleinberg. 2002. Bursty and hierarchical structure in streams.

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

[Ku et al.2006] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 100–107.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1).

[Mei et al.2007] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM.

[Mukherjee and Liu2012] Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 339–348. Association for Computational Linguistics.

[Owoputi et al.2013] Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

[Patino et al.2012] Anthony Patino, Dennis A Pitta, and Ralph Quinones. 2012. Social media's emerging importance in market research. *Journal of Consumer Marketing*, 29(3):233–237.

[Popescu and Etzioni2007] Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pages 9–28. Springer.

[Popescu and Jain2011] Ana-Maria Popescu and Alpa Jain. 2011. Understanding the functions of business accounts on twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 107–108. ACM.

[Sakai2007] Tetsuya Sakai. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information Processing & Management*, 43(2):531–548.

[Schwarz1978] Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Spina et al.2012] Damiano Spina, Edgar Meij, Maarten de Rijke, Andrei Oghina, Minh Thuong Bui, and Mathias Breuss. 2012. Identifying entity aspects in microblog posts. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1089–1090. ACM.

[Titov and McDonald2008] Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

[url2013] 2013. `phys.org/news/2013-03-global-unexpected-shipping-routes-arctic.html`.

[url2014] 2014. `http://cran.r-project.org/web/packages/bursts/index.html`.

[Zhao et al.2011] Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee Peng LIM, and Xiaoming Li. 2011. Topical keyphrase extraction from twitter. In *ACL*. ACM.