Entity-Centric Topic-Oriented Opinion Summarization in Twitter

Xinfan Meng[‡][*] Furu Wei[†] Xiaohua Liu[†] Ming Zhou[†] Sujian Li[‡] Houfeng Wang[‡]
[‡]MOE Key Lab of Computational Linguistics, Peking University
[†]Microsoft Research Asia
[‡]{mxf, lisujian, wanghf}@pku.edu.cn
[†]{fuwei,xiaoliu,mingzhou}@microsoft.com

ABSTRACT

Microblogging services, such as Twitter, have become popular channels for people to express their opinions towards a broad range of topics. Twitter generates a huge volume of instant messages (i.e. tweets) carrying users' sentiments and attitudes every minute, which both necessitates automatic opinion summarization and poses great challenges to the summarization system. In this paper, we study the problem of opinion summarization for entities, such as celebrities and brands, in Twitter. We propose an entity-centric topic-based opinion summarization framework, which aims to produce opinion summaries in accordance with topics and remarkably emphasizing the insight behind the opinions. To this end, we first mine topics from #hashtags, the human-annotated semantic tags in tweets. We integrate the #hashtags as weakly supervised information into topic modeling algorithms to obtain better interpretation and representation for calculating the similarity among them, and adopt Affinity Propagation algorithm to group #hashtags into coherent topics. Subsequently, we use templates generalized from paraphrasing to identify tweets with deep insights, which reveal reasons, express demands or reflect viewpoints. Afterwards, we develop a target (i.e. entity) dependent sentiment classification approach to identifying the opinion towards a given target (i.e. entity) of tweets. Finally, the opinion summary is generated through integrating information from dimensions of topic, opinion and insight, as well as other factors (e.g. topic relevancy, redundancy and language styles) in an unified optimization framework. We conduct extensive experiments on a real-life data set to evaluate the performance of individual opinion summarization modules as well as the quality of the produced summary. The promising experiment results show the effectiveness of the proposed framework and algorithms.

[*]Contribution during internship at Microsoft Research Asia.

KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$15.00.

Categories and Subject Descriptors

I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Experimentation

Keywords

Opinion summarization, sentiment analysis, topic analysis, #hashtag, Twitter

1. INTRODUCTION

Twitter[1] is a microblogging service that allows people to publish short messages (i.e. tweets) up to 140 characters to share with others what is happening, what is interesting, as well as their feelings and opinions. In the recent several years, Twitter has become an extremely popular site. It has been reported that about 200 million[2] tweets are posted on June 30, 2011. In these tweets, people not only share their daily update information or personal conversation, but also exchange their opinions towards a broad range of topics. For example, they may talk about the newly released electronic products, express their feelings to celebrities, and deliver their comments on politicians or global events, etc. Consequently, we can use Twitter to survey public opinions. In this paper, we are particularly interested in the sentiment towards certain entities , such as celebrities, brands and products. These opinions are extremely useful in many real-life applications. For instance, politicians can study opinions in Twitter in order to know their public image, and company can study opinions in Twitter to obtain customer feedbacks.

The huge volume of tweets carrying sentiments necessitates automatic summarization techniques, which assists users in consuming these opinions. Ideally, we expect the opinion summary to convey the overview of the opinions for a specific entity from the public perspective. However, people may express opinions towards different aspects, or topics, of an entity. Take the US president Obama as an example, people may criticize (in a negative manner) his economic revitalization plans, but firmly support (in a positive attitude) Obama's counter-terrorism polices. This suggests a fine-grained summary, in terms of the different topics, instead of the general summary that mixes up the opinions

[1] http://twitter.com

[2] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

from all aspects regarding Obama. Furthermore, in Twitter, many tweets simply reveal the sentiment information. For example, "*I love Obama! #libya*". These tweets certainly account for accumulating the overall sentiment for the target, but do not deliver much information for the underlying insight behind the opinions, which are even more important for users. On the other hand, there are tweets in Twitter that explicitly state the factors leading to the positive or negative opinions. We show a typical example: "*I think it is time to say that Barack Obama deserves credit for backing up his words with action on #Libya despite domestic opposition.*" We can clearly understand the reasons underlying the sentiment from the tweet. To summarize, the opinion summary should include three key elements, which are **topic**, **sentiment** and **insight**.

In this paper, we study the problem of opinion summarization for entities in Twitter. In particular, we propose an entity-centric topic-based opinion summarization framework, which aims to produce opinion summaries in accordance with topics and remarkably emphasizing the insight behind the opinions. It has been well-recognized that tweets are short, noisy and ungrammatical, which issues tough challenges for topic analysis, sentiment analysis as well as identifying tweets with insights in Twitter. As for topics analysis, we leverage the human-annotated semantic tags, i.e. #hashtags to assist identifying topics from tweets. #hashtags are created organically by Twitter users as a way to categorize messages and to highlight topics, which is done by simply prefixing a word or a phrase with a symbol # to mark keywords or topics in tweets, such as "#hashtag". For example, in the tweet "*Neelie Kroes, Vice President of European Commission, Digital Agenda Commissioner, will be Keynote speaker at #www2012 : http://www2012.org/*", the author marked up the topic keyword "#www2012" as a #hashtag. Besides, in Twitter, people often group several words together to form a #hashtag, for example, #IraqWar and #HealthCare. Those #hashtags have clear boundaries and often can be viewed as semantic concepts, hence are more suitable as topics than n-gram phrases. The extensive use of #hashtags makes Twitter more expressive and welcomed by people. We measure on a data set with around 0.2 million tweets and find that around 23% of them have at least one #hashtag in each. Among all unique #hashtags, we sample 100 #hashtags and find 75% are useful for topic extraction. The statistics show a great potential for mining topics from #hashtags in Twitter.

Consequently, we integrate #hashtags as weakly supervised information into topic modeling algorithms to obtain better interpretation and representation for calculating the similarity among them, and then adopt Affinity Propagation algorithm to group #hashtags into coherent topics. We develop a target (i.e. entity) dependent sentiment classification approach to identifying the opinion towards a given target (i.e. entity) of tweets. We incorporate the dependency relationship between the sentiment lexicons and the target (i.e. entity) into the lexicon based sentiment classification framework. Subsequently, we use template methods to identify insightful tweets that revealing reasons, expressing demands or reflecting viewpoints to embed deep insights into the opinion summary. However, template methods may suffer from the problem of low recall, we thus employ paraphrasing approach to conduct template expansion. Finally, the opinion summary is generated through integrating information from the above-mentioned dimensions as well as other factors (e.g. redundancy and language styles) in an unified optimization framework.

We conduct extensive experiments on a real-life data set to evaluate the performance of individual opinion summarization modules as well as the quality of the produced summary. The promising experiment results show the effectiveness of the proposed framework and algorithms. To the best of our knowledge, the work in this paper is the first attempt on opinion summarization in Twitter.

2. RELATED WORK

Opinion summarization is a broad and diversified research topic. Most existing work focuses on summarizing opinions from user generated content such as product reviews [11, 12], movie reviews [28] or hotel reviews [26]. They follow the aspect-based opinion summarization paradigm which analyzes sentiment on fine-grained features or aspects of a product. The features can be the functionality of the product, like camera functionality of a mobile phone, or a part of the product, like the screen of a mobile phone. In these work, aspects are extracted from reviews by association rule mining [11, 12] or aspect-sentiment topic models [18, 25, 19], extensions of topic models. Then, the opinion towards each aspect is extracted and summarized. In the simplest case, the summary is a positive or negative label aggregated over the user generated content. Different methods are proposed to assign the sentiment labels for aspects. Some of them leverage existing sentiment classifier, while others directly model aspect-sentiment relation using the aspect-sentiment models [19]. The interested readers are referred to [22] and [16] for a general introduction to opinion summarization, and [14] for a more recent survey on opinion summarization. The above-mentioned work can be viewed as opinion summarization on restricted domain (product reviews, movie reviews, etc.), unlike our approach, which is applied in Twitter, a general domain opinion source.

Sentiment analysis on microblogging services like Twitter is also receiving popularity. Barbosa and Feng conduct sentiment classification on tweets via two-stage SVM classifier [1]. They focus on selecting features and combining different label sources in order to remove noises in Twitter. Davidov et al. incorporate #hashtags and smileys from tweets as sentiment labels [8]. O'Connor et al. determine the sentiments by subjective Lexicon [21] . Jiang et al. study target-dependent Twitter sentiment classification [13]. Wang et al. [27] propose to conduct sentiment classification on #hashtags, which are coarsely regarded as topics in their work. These work mainly focuses on sentiment classification, which apparently differs from the work described in this paper.

3. ENTITY-CENTRIC TOPIC-ORIENTED OPINION SUMMARIZATION

We start this section by a formal definition of the task of entity-centric topic-oriented opinion summarization in Twitter and then present an overview of the proposed approach.

Given a set of tweets $\mathcal{T}$ mentioning an entity e, we aim to produce a opinion summary $O = \{O_1, O_2, \ldots, O_N\}$, where,

- N is the number of topics;
- $O_i, i \in [1, N]$, is a $< label, \{t\} >$ pair, where *label* is

the label of the topic, and $\{t\} \subseteq \mathcal{T}$ is a set of insightful opinionated tweets.

We refer a tweet $t_i \in \{t\}$ as an *insightful opinionated tweet* to indicate a tweet not only conveys opinions but also provides insight. We show an example of insightful opinionated tweet as follows.

> I think it is time to say that Barack Obama deserves credit for backing up his words with action on #Libya despite domestic opposition.

The author not only expresses his/her positive attitude towards President Obama, but also explicitly states that this attitude is the result of Obama's actions on Libya. This type of insightful opinionated tweets conveys the insight behind the opinions, which is especially important for opinion summarization.

Accordingly, we divide the system into two main parts, namely topic extraction and opinion summarization. Firstly, we extract topics for an entity from tweets containing the entity. Secondly, we identify the opinionated tweets with insight information to compose the opinion summary and organize the summary in accordance with topic extraction results.

4. TOPIC EXTRACTION

In this paper, we propose to mine topics from #hashtags. #hashtags are human annotated tags for providing additional context and metadata to tweets, which are used to categorize messages and to highlight topics. We use #hashtags as candidate topics. We conduct a comprehensive study on #hashtags and categorize #hashtags according to their usage in Table 1. In this paper, we focus on **category keywords**, **entities** as well as **events and issues**. We develop a rule-based classifier to identify the #hashtags of these types. Specifically, for category keywords #hashtags, we collect a category dictionary from the Open Directory Project (http://www.dmoz.org/). For person/location #hashtags, we collect a dictionary from Freebase (http://www.freebase.com/). For events/issues #hashtags, we first use a bi-gram language model segmenter[3] to split #hashtags into multiple words. And then we check if some of these words are in our person/location dictionary. If one or more words are in our dictionary, this #hashtag tends to be an event or issue, #LondonRiot and #OccupyChicago for instance. For the remaining #hashtags, we use *tagness*[4] to determine if they are candidate topics. Tagness for a #hashtag is defined as the occurrences of this #hashtag divided by the total occurrences of its content, i.e. without # symbol. For #hashtag may contain multiple words, when computing tagness, we also segment the words and count the occurrences of the segmented word sequences. When the tagness of a #hashtag is smaller than a threshold we set empirically(0.85 in this paper), it will be one of our candidates topic #hashtags. With tagness measurement, we can remove many #hashtags in the other 4 categories, such as #fb and #tcot, since they are always used as #hashtags.

4.1 Graph-based Topic Extraction

In this paper, we model the task of extracting topics as clustering #hashtags into coherent groups. Particularly, we create a weighted undirected graph $G =< \mathcal{H}, \mathcal{E}, f >$. $\mathcal{H}$ is the node set, and each node in the graph is a #hashtag; $\mathcal{E}$ is the edge set, and $f(e = (h_i, h_j) \in \mathcal{E}) \rightarrow \mathcal{R}$, $1 \leq i, j \leq |H|$, is the weight function to measure the relatedness between the two nodes (i.e. #hashtags), h_i and h_j. Then, we run Affinity Propagation [9], a state-of-the-art clustering algorithm on G. The input of the Affinity Propagation clustering algorithm is the #hashtags pairwise relatedness matrix, and the output are the #hashtags clusters and the centroids of clusters.

As for the #hashtags relatedness, we consider three kinds of metrics,namely co-occurrences, context similarity and distributional similarity via weakly supervised topic models.

Co-occurrences Relation.

$$f((h_i, h_j) \in e) = \begin{cases} 1 & \text{if } h_i \text{ and } h_j \text{ co-occur in a tweet} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Context Similarity.

$$f((h_i, h_j) \in e) = \frac{\vec{v}(h_i) \bullet \vec{v}(h_j)}{|\vec{v}(h_i)| \times |\vec{v}(h_j)|} \tag{2}$$

$\vec{v}(h)$ is the centroid context vectors of #hashtags h, which can be obtained from the output of the Affinity Propagation algorithm.

Topic-Aware Distributional Similarity.

Co-occurrences metric discards words that are not #hashtags; context similarity metric assigns equal importance to every word and every #hashtag occurred in a tweet. In order to obtain better representation and interpretation of #hashtags, we use Labeled LDA [23] to learn #hashtags-words correspondence, which associates individual word in a tweet to appropriate #hashtags. We model the #hashtag(s) in a tweet as the label(s) of the tweet and other words as the document for Labeled LDA. We can thus obtain the #hashtag-word distribution, which is similar to topic-word distribution in LDA. Then, we use negative symmetric KL divergence of distribution over words as the similarity between #hashtags. We use Labeled LDA instead of supervised topic model [3], since Labeled LDA can handle multiple-label cases.

$$f((h_i, h_j) \in e) = -\frac{1}{2}\Big[KL\big(p(W_{h_i})||p(W_{h_j})\big) + KL\big(p(W_{h_i})||p(W_{h_j})\big)\Big] \tag{3}$$

where $p(W_{h_i})$ and $p(W_{h_j})$ are the word distributions of h_i and h_j, respectively.

4.2 Topic Labeling and Assignment

Once we obtain the topics from the previous step, we label each topic with a representative #hashtag. We use the topic centroids output by Affinity Propagation as the topic labels. We assign one or more suitable topics extracted to each tweet. Firstly, for a tweet with #hashtag(s), we assign it the topic(s) corresponding to every #hashtag in the tweet. Hence we may assign multiple topics to a tweet. Secondly, for a tweet without #hashtags, we predict its topic using a SVM classifier trained on tweets that have been assigned topics in the first step. We use bag-of-words features for the classifier.

[3] http://norvig.com/ngrams/

[4] http://energy.twex.poeschko.com/metrics/

Category	Examples	Description
Category keywords	#sports, #economy	General category keywords
Person and location	#Bush, #Libya	People or locations involved
Events and Issues	#LibyaWar, #OccupyWallStreet	Describing events and issues, usually with multiple words
Content clues	#NowPlaying, #WhenIWasAKid	Clues or hints to the tweets content
Sentiment	#fun, #fail	Describing the sentiment expressed in the tweet
Source and media	#fb, #NYT	The sources of the tweets.
Discussion channel	#tcot #tlot	Forwarding the tweets to the related discussion groups

Table 1: #Hashtag categorization

5. OPINION SUMMARIZATION

We use the insightful opinionated tweets to compose the opinion summary. In the following sections, we first describe insightful and opinionated tweets classification and then present an optimization framework to integrate **topic**, **sentiment** and **insight** to generate the opinion summary.

5.1 Insightful Tweet Classification

It is difficult to come to an universal and formal definition of insightful tweets. For our application, we define 7 types of insightful tweets and list them in Table 2, based on the reason that all these types of tweets provide deeper insights and are more useful to the readers.

As a baseline, we use a binary classification approach to select the insightful tweets. In particular, we use LibSVM [5] with bag-of-words features and linear kernel. However, based on our observation of the tweets dataset, only about 10% of the tweets can be regarded as insightful opinionated tweets, which indicates this is an imbalance dataset. In order to alleviate the data imbalance problem, we increase the weight of positive instance [2], but the classifier's performance does not become better.

Alternatively, we can use the patterns defined in Table 2 to identify insightful tweet. A naive pattern matching method is to manually create a pattern list and match the pattern strings against the candidate tweets, and then we predict the matched tweets as insightful. However, this straightforward pattern matching method causes two problems. First, given a tweet containing multiple entities, it makes the same prediction for different entities, irrespective of the syntactic relations between the pattern and the entities. Second, the coverage of manually created pattern list is low.

Therefore, instead of straightforward pattern matching, we use a syntactic-constrained pattern matching approach. We first use Stanford Parser[5] to parse pattern phrases to obtain pattern syntax trees. And then we parse tweets to obtain tweets syntax trees. Finally we match the pattern syntax trees against the pattern syntax trees. To efficiently create a high coverage pattern set without hurting precision too much, we adopt an automatic pattern generalization paradigm. We first define an initial set of patterns and then generate new patterns for each pattern in the initial pattern set. For example, given a pattern "that is why", we generate two new patterns "which is why" and "this is why" and they share the same meaning. To avoid generating grammatically incompatible patterns, we also enforce the syntax constraint, i.e. the pattern generated should have the same syntax tree as the original, in order to ensure the pattern and the new pattern(s) mutually substitutable in the syntax tree. Therefore we use a paraphrase generation algorithm [4], which output patterns having the same syntactic structures to the input patterns by applying syntactic constraints to the underlying phrase extractions and paraphrasing substitution approaches, to generate new patterns.

5.2 Opinionated Tweet Classification

The opinion summary consists of opinionated tweets, i.e. tweets containing positive or negative sentiment regarding the entity. We thus need an entity (also called target) dependent sentiment classifier to identify the sentiment orientation (positive (P), negative (N) or neutral (O)) of a tweet. In this paper we build the tweet-level sentiment classifier based on sentiment lexicon.

Our lexicon-based sentiment classifier relies on sentiment dictionary matching, or in other words, sentiment lexicon words counting. Specifically, given a sentiment lexicon $\mathcal{D}$, each word in $\mathcal{D}$ is annotated with a prior sentiment orientation, either positive or negative. The lexicon based approach counts the occurrences of the positive (c_p) and negative (c_n) words and determines the sentiment orientation of a sentence (s) simply by aggregating c_p and c_n. However, many named entities contain sentiment words. As a result, we conduct named entity recognition (NER) on sentence, and sentiment words contained in any entity will not be counted.

$$SO(s) = \begin{cases} P & \text{if} c_p > c_n \\ N & \text{if} c_p < c_n \\ O & \text{otherwise.} \end{cases} \tag{4}$$

Here, $SO(s)$ denotes the sentiment orientations of the tweet s. We should pay additional attention to the negation phenomenon in lexicon-based sentiment analysis. Negation expressions in sentences might cause sentiment negation, i.e. reverse the sentiment orientation to its opposite side. Basically, we adopt the categorization for negation as introduced in [7], which categorizes the negation words into two classes: content negation words (such as "not", "never") and function negation words (such as "eliminate", "reduce"). According to both [20] and [7], to incorporate the two types of negators will benefit the accuracy of sentiment classification greatly. We take a simple and heuristic approach to tackling the negation problem in our lexicon-based sentiment analysis. We invert the local sentiment orientation of a sentiment word $w \in \mathcal{D}$ to its opposite orientation whenever a negation expression *neg* is found preceding w and the distance in words between *neg* and w is smaller than a predefined threshold (5 in our experiments).

It should be emphasized that we are working on target-dependent sentiment classification. For this purpose, before we count the positive and negative sentiment words, we conduct a classification to determine whether the sentiment word (w) is used to depict the target (e). This is achieved

[5] http://nlp.stanford.edu/software/lex-parser.shtml

Category	Example	Example Patterns
Reasoning tweets	"And that is why I voted for President Obama. No more war!"	that is why, because, because of, owing to etc.
Appeal tweets	"Please re-elect President Obama. Dont let these rep. And other blind you with the truth."	I hope, stop, please, etc.
Causal tweets	"Obamacare was, is, and always will be a fatally flawed idea that will only lead to bankruptcy."	lead to, contribute to, result in , etc.
Comparative tweets	"Pinocchio is looking like an honest guy compared to the GOP candidates."	comparing to, compared with, etc.
Opposite opinionated tweets	"@BarackObama I think you are a remarkable man but I come from a family of republicans So I'll be supporting the Republican candidacy"	anyway, however, etc.
Viewpoint tweets	"I wonder how many followers Obama will have after he leaves office and we have another president ? "	I believe, I think, I wonder, I bet, we need to, etc.
Subjunctive mood tweets	"Martin Luther King Jr. would have supported Occupy Wall Street"	would be, would have, etc.

Table 2: Examples of insightful tweets

by a binary SVM classifier which predicts whether the sentiment of e should be dependent on w or not. We design a set of features to facilitate the classification as follows.

- The distance in word between w and e;
- Whether there are other entities between w and e;
- Whether there are punctuation(s) between w and e;
- Whether there are other sentiment word(s) between w and e;
- The relative position of w and e: w is before or after e;
- Whether these is a dependency relation between w and e. We conduct a dependency parsing on the tweet s using the MST Parser[6], and check whether there is dependency link between e and w. We do not consider the dependency relation labels between e and w.

For a given entity e, we only consider the sentiment words w which are classified as dependent to e in the lexicon based sentiment analysis approach as described in Equation 4.

5.3 Optimization-based Summary Generation

By using the techniques described in previous sections, we can already obtain insightful opinionated tweets for each topic towards the given entity. However, these tweets can not be directly output as the opinion summary. First, Many of them are difficult to understand. Unlike news report and other formal text, people use slangs, abbreviations and emoticons extensively in Twitter. Tweets containing too many informal symbols and words are unreadable and should not be included in the opinion summary. Second, they contain highly redundant information, since retweeting is popular in Twitter, which make many tweets extremely similar or even identical.

Consequently, we use an optimization framework to select the most relevant, representative and readable tweets as the final opinion summary. The framework considers three factors simultaneously, namely, relevancy, redundancy and readability of tweets. Intuitively, we model this problem as selecting a subset of tweets from all tweets, so as to minimize the information loss in discarding other tweets. Formally, we define the tweets selection problem as selecting a subset of tweets P from tweet set T_k for topic k to

$$\min_{P \subset T_k} \sum_{t \in T_k - P, p \in P} \left(\operatorname*{argmin}_{p \in P} D(t,p) \right)$$

where $D(t,p)$ is the cost function of representing tweet t with tweet p. This is an NP-hard problem in general, but POPSTAR [6, 24] algorithm can be used to solve this problem approximately. We define the cost function for opinion summarization as

$$D_k(t_i, t_j) = L(t_i) * T_k(t_i) * Re(t_i, t_j)$$

where t_i and t_j are two tweets and k is the topic index. $L(t_i)$ is the language style score of t_i. We favor the more readable tweet.

$$L(t_i) = 1 + \frac{\text{\# of words out of vocabulary}}{\text{tweet length}}$$

$T(t_i)$ is the topic relevance score of t_i, defined as KL divergence between term distribution of t_i and topic label l_k.

$$T(t_i) = KL(t_i, l_k)$$

$Re(t_i, t_j)$ is the redundancy score between tweet t_i and t_j, defined as KL divergence between word distribution of t_i and t_j.

$$Re(t_i, t_j) = KL(t_i, t_j)$$

The final output from our optimization framework are several representative opinionated insightful tweets per topic. Note that we use a asymmetrical cost function here, $D_k(t_i, t_j) \neq D_k(t_j, t_i)$. As a result, the optimization procedure will automatically select the direction with less cost, hence it tends to keep the more readable and relevant tweets.

6. EXPERIMENTAL STUDY

In this section, we present the evaluation results on a real-life data set. We first provide thorough experiments on each component, i.e. topic extraction, insightful tweet classification as well as opinionated tweet classification, and then we present the evaluation results on the final opinion summaries.

[6] http://www.ryanmcd.com/MSTParser/MSTParser.html

6.1 Data

We collect the evaluation data set from 15th September 2011 to 20th October 2011 using the Twitter API[7]. We mainly focus on two types of entities, people (e.g. politicians and celebrities) and brands (e.g. electronic devices company and software company). We select 6 entities, 3 for people and 3 for brands. The entities include, "Obama", "Lady Gaga", "David Cameron", "Microsoft", "Apple", and "Nokia". We use these entities as queries and continuingly crawled tweet mentioning them to build the evaluation corpus. We do not use entity normalization techniques here because the amount of tweets is enormous and we can collect enough data simply through string matching. We collected 201, 234 tweets in total. Table 3 shows the statistics of our data set.

Person	#tweets	Brand	#tweets
Obama	34,980	Apple	39,208
Lady Gaga	43,992	Microsoft	30,993
David Cameron	21,782	Nokia	30,279
Total	100,754	Total	100,480

Table 3: Descriptions of the evaluation corpus

As a preliminary study, we investigate the effectiveness of using #hashtags as topics. We count the number and the percentage of tweets containing zero, one or more #hashtags, and we find out that about 23% of tweets contain #hashtags, and 8% of tweets contain more than 1 #hashtags. These #hashtags can cover a large proportion of the topics in Twitter.

Furthermore, we examine the feasibility of using #hashtags to extract topics. We obtain 15,865 unique #hashtags and 83, 713 #hashtags from the evaluation corpus. Then, we sample 100 #hashtags from the set of the 15,865 #hashtags and manually examine the tags. We discover that 75 #hashtags can be used to represent topics, which supports our proposal of extracting topics from #hashtags.

6.2 Evaluation of Topic Extraction

The aim of the experiments in this section is to evaluate the effectiveness of topic extraction from tweets using the graph-based methods. In particular, we evaluate the performance of topic clustering and topic assignment.

The evaluation of topic extraction is challenging because it is difficult to collect the "gold standard" data set. Though it is possible to identify and label topics from tweets manually, the time cost is prohibitive. Therefore we conduct our evaluation by comparing the topics generated by our methods and baselines.

6.2.1 Evaluation of Topic Clustering

Here we evaluate the effectiveness of topic clustering. The aim is to assess the coherency of the produced #hashtags topics. We conduct the tokenization on the evaluation corpus with ark-tweet-nlp [10], and then remove stopwords, numbers (1.2, 100, $5 etc.), URL and words starting with "@" (accounts in Twitter). Finally we use Porter stemmer[8] to stem the words and #hashtags.

We run the Labeled LDA implementation in the Stanford Topic Modeling Toolbox[9] with the default settings to obtain the #hashtag-word distribution from the preprocessed evaluation corpus. Then we feed the distribution into the Affinity Propagation algorithm[10] and produce the #hashtags clusters.

We evaluate the clustering result by purity [17]. Table 4 shows the result. From the table, we can clearly find that Co-occurrence and Cosine method produce more clusters and smaller clusters than the Labeled LDA method. This indicates that Co-occurrence method and Cosine method can not effectively model the relation between #hashtags and consequently are not capable of finding strong relation information to group similar #hashtags. Moreover, the distributional similarity approach (Labeled LDA) can greatly improve the performance.

6.2.2 Evaluation of Topic Assignment

We evaluate the effectiveness of our topic assignment approach in this section. We sample 100 tweets for 3 topics per entity as the evaluation data set, and collect a corpus of 1,800 tweets in total.

We report the topic assignment accuracy, the number of tweets correctly assigned divided by the total number of tweets, in Table 5. The topics are generated from the Label LDA approach, which gives the best topic extraction results as we have stated in Section 6.2.1. We notice that the accuracy for the singer Lady Gaga is substantially lower than that of politicians. Examining the dataset, we find that the tweets about Lady Gaga contain more noises.

6.3 Evaluation of Opinion Summarization

In this section, we present the evaluation results of opinion summarization. First, we evaluate the performance of identifying insightful tweets and opinionated tweet classification. Then, we assess the quality of opinion summary.

6.3.1 Evaluation of Insightful Tweets Classification

To evaluate the effectiveness of insightful tweets classification, we annotate 300 tweets per entity and obtain a test set of 1,800 tweets.

We conduct experiment on both the SVM baseline and the pattern matching approach. We train the SVM baseline with 100 tweets as training data (14 are insightful tweets). We experiment two setting of our pattern approach. In the first setting, we pick several pattern phrases for each type of insightful tweets and in the end we collect a pattern set of 20 patterns phrases. In the second setting, we use the same pattern phrase set but use the syntactic-constrained paraphrase method to conduct pattern expansion. After pattern expansion, our pattern set expands from 20 to 187. We evaluate the performance of pattern approach without generalization (Pattern_{man}) and pattern approach with generalization (Pattern_{gen}) respectively. We report the performance of these three methods in Table 6. The precision of baseline classification is very low, owing to the data imbalance problem. Meanwhile, patterns selected by human tend to have high precision but very low recall. After pattern expansion, recall increases by 13% while precision only decreases by 4%. This indicates that our pattern expansion method is effective.

[7] https://dev.twitter.com/

[8] http://tartarus.org/martin/PorterStemmer/

[9] http://nlp.stanford.edu/software/tmt/tmt-0.3/

[10] http://scikit-learn.sourceforge.net

	Methods	Obama	Lady Gaga	David Cameron	Microsoft	Apple	Nokia	Average
# of clusters	Labeled LDA	12	16	13	15	19	17	15
	Cosine	35	33	27	17	24	25	27
	Co-occurrence	36	34	19	26	23	34	29
Accuracy	Labeled LDA	0.63	0.54	0.55	0.73	0.65	0.68	0.63
	Cosine	0.35	0.43	0.37	0.47	0.52	0.47	0.435
	Co-occurrence	0.33	0.44	0.48	0.46	0.45	0.41	0.428

Table 4: Performance of #hashtags clustering

Methods	Obama	Lady Gaga	David Cameron	Microsoft	Apple	Nokia	Average
Labeled LDA	76.7%	66.7%	73.3%	83.7%	75.0%	73.7%	75.7%

Table 5: Topic assignment accuracy

Method	Precision	Recall	F1
SVM Baseline	28.0%	46.5%	35.0%
Pattern$_{man}$	76.0%	37.0%	49.8%
Pattern$_{gen}$	72.0%	49.8%	58.9%

Table 6: Performance of insightful tweets classification

Precision	Recall	F1-Score
86.5%	80.6%	83.4%

Table 7: Performance of the target-lexicon dependency classification

Classifier	Precision	Recall	F1
Baseline	62.6%	65%	61.7%
System	85.3%	61%	71.1%

Table 8: Performance of the opinionated tweet classification

6.3.2 Evaluation of Opinionated Tweet Classification

In this section, we evaluate the performance of opinionated tweet classification as described in Section 5.2. We also present the evaluation result for the dependency classification between target and sentiment word.

In our experiments, we employ the sentiment lexicon from the General Inquirer[11]. We manually annotate 500 tweets which are sampled from the tweets containing the 6 entities as mentioned in Section 6.1. We ask two annotators to conduct the annotation. Each annotator is asked to annotate each tweet as positive, negative or neutral, given the specific target (entity). We also ask the annotators to annotate the dependency between the sentiment word and the target in the tweet. Finally, we obtain 420 tweets, each of which is annotated with the same sentiment label by the two annotators. This indicates the inter-agreement of the annotators is 84%. We use the 420 tweets as our evaluation set for opinionated tweets classification. The evaluation set contains 203 positive tweets, 126 negative tweets and 91 neutral tweets.

Table 7 shows the precision, recall, and F1-Score on 5-fold cross valuation of the target-lexicon dependency classification. As we can see, it yields very encouraging results.

Table 8 shows the results for target-dependent tweet-level sentiment analysis results. Since the evaluation data set is also used in the target-lexicon dependency classification, we report the results using the same 5-fold cross validation strategy here. The system and baseline method in Table 8 indicates whether we incorporate the target-lexicon dependency classification into the lexicon based sentiment classification approach or not. As can be seen from the table, the target dependent sentiment classification can greatly improve the performance. Especially, the precision has been significantly improved, which is very important in the context of our work because there will always be huge amount of tweets available and thus we favor precision over recall.

6.3.3 Evaluation of Summary Generation

Here we evaluate the effectiveness of our method to generate opinion summary in Twitter. To build the evaluation data set, we assume that the topics given by the topic extraction procedure are accurate and select 3 topics for each entity, and then for each topic we select 100 tweets as the corpus for opinion summarization. Then we manually select 5 insightful opinionated tweets from those 100 tweets for each topic as the "ground truth" opinion summary of that topic. Totally we have an evaluation set of 18 opinion summaries on 1800 tweets.

We adopt ROUGE-SU4 measure to evaluate the final summary produced [15]. For comparison, we also experiment a setting of our approach which ignores the language style, i.e. $L(t_i)$ always equals to 1. We compute the average precision, recall and F1-Score of 3 summaries of each entity and report results for both settings in Table 9. The results indicate that employing linguistic knowledge improves the opinion summary quality. We conclude that language style is vital for opinion summarization in Twitter.

7. CASE STUDY

To provide deeper insight into the advantages of topic-oriented insight-based opinion summary, we use the opinion summary towards "Obama" as a case study. Table 10 shows the topic clusters extracted from the tweets, along with their topic labels. We select the #hashtags by its frequency in the clusters. From this table, we observe that the topic labels directly suggest the topics, and other #hashtags in the cluster are also tightly related to the topic. For example, the label of the first topic is "#occupywallst", which immediately suggests that it is a topic about the OccupyWallStreet movement. And other words in this cluster provide fur-

[11] http://www.wjh.harvard.edu/∼ inquirer/

Methods	ROUGE-SU4	Obama	Lady Gaga	David Cameron	Microsoft	Apple	Nokia	Average
Our approach	Precision	0.2392	0.2443	0.2876	0.2764	0.2812	0.2573	0.2643
	Recall	0.2488	0.2132	0.2245	0.2712	0.2324	0.2319	0.2370
	F1	0.2439	0.2277	0.2522	0.2738	0.2545	0.2439	0.2500
-Language	Precision	0.1537	0.1824	0.2128	0.2412	0.1932	0.1754	0.1931
	Recall	0.1871	0.2026	0.2201	0.2207	0.1538	0.1928	0.1961
	F1	0.1688	0.1920	0.2164	0.2305	0.1713	0.1837	0.1946

Table 9: Performance of summary generation

Topic label	Topic
#occupywallst	#wallstreet #occupydenv #fireobama
#iran	#middleeast #israel #nuclear #islam
#obama2012	#hermancain #rickperri #energi #vote

Table 10: Topic cluster for "Obama" with our method

Topic	Top ranking words ranked by $p(words|topic)$
1	wall street occupi protest support parti ows
2	iran order execut withdraw sign trade plot
3	support elect becaus last fast reelect work

Table 11: Topic cluster for "Obama" with LDA

ther information such as the event location (#wallstreet), the slogan (#fireobama). The topics extracted are easy to interpret and understand.

For a comparison with popular topic extraction methods, we also run LDA on the same dataset and generate Table 11 with K = 12, the same as the number of topics found by our method. We use the LDA implementation in Stanford Topic Modeling Toolbox with the default setting. From the clusters we select 3 example clusters that describe approximately the same topic produced by our method. Topic 1 conveys the same amount of information with more words. In particular, they describe the location with two words, "wall" and "street". Topic 2 in LDA contains many verbs, while the corresponding topic 2 in our approach just contain nouns and/or noun phrases and are easier to interpret. Comparison of Topic 3 demonstrates superior performance of our approach. Topic 3 in LDA contains some common words such as "becaus", "last" and "fast", while our approach select readable topic #hashtags. Moreover, our approach directly provides a topic label to represent the main idea of the topic, which is not available when using LDA.

In Table 12, we present the opinion summaries, consisting of insightful opinionated tweets for each topic. This summary is a tight integration of topic, sentiment and insight. The patterns and clue words for identifying the insightful tweets are highlighted using the bold font. We select 4 representative topics for Obama, namely Libya issue, OccupyWallStreet, HealthCare and jobs plan, covering both domestic and foreign affairs. We can gain many insights from this table. First, the 4 topic labels correctly and concisely summarize the 4 hot debated topics regarding president Obama. Second, on the topic of Libya issue, Obama is praised for saving American soldiers' lives for his intervention policy, though at the same time he is disputed on how he deals with the violence issue in Libya and United States. Third, for the other three topics, Obama is facing criticisms for his ineffective policy. Finally, it can be concluded that Obama receives more supports on foreign affairs than on domestic affairs. We also discover that people tend to provide more insights behind the opinions in their negative tweets.

From this case study, we see that the integration of topic, sentiment and insight provide a powerful way for summarizing opinions regarding entities.

8. CONCLUSION

In this paper, we present an entity-centric topic-oriented opinion summarization framework, which is capable of producing opinion summaries in accordance with topics and remarkably emphasizing the insight behind the opinions in Twitter. We decompose the opinion summarization into three dimensions, namely topic, opinion and insight; the opinion summary is generated by integrating these three dimensions as well as other factors (e.g. redundancy and language styles) in an unified optimization framework. We develop a template based method to effectively identify the insightful tweets with high precision; we use target dependent sentiment classification to identify opinionated tweets regarding the entities. Extensive experiments are conducted to evaluate the performance of both the individual summarization components and the overall summarization results. We also present a case study of the produced opinion summary, which further demonstrates the effectiveness of the proposed opinion summarization framework and algorithms.

In the future, we will further study the semantics underlying #hashtags, which we can make use of to extract more comprehensive and interesting topics. We believe topic based opinions summarization will benefit from the deeper understanding of #hashtags.

9. ACKNOWLEDGMENT

This research was partly supported by National High Technology Research and Development Program of China (863 Program) (No. 2012AA011101) and National Natural Science Foundation of China (No.91024009, No.60973053)

10. REFERENCES

[1] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proc. COLING*, pages 36–44, 2010.

[2] A. Ben-Hur and J. Weston. A user's guide to support vector machines. *Methods in Molecular Biology*, 609:223–239, 2010.

[3] D. M. Blei and J. McAuliffe. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128, 2008.

[4] C. Callison-Burch. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proc. EMNLP*, October 2008.

Topic label	Opinion Summaries	Sentiment
#libya	**I think** it is time to say that Barack Obama deserves credit for backing up his words with action on #Libya despite domestic opposition.	Positive
	Thanks to Obama's approach to Libya, not 1 American soldier was lost.	Positive
	Film **comparing** Obama's praise of public protests in Libya/Syria **with** the violence of arrests on Occupy Wall St	Negative
#occupywallst	@BarackObama I'm running out of hope. **Please** replace Geithner w/ Reich or Krugman #wallstreetoutofwhitehouse #OWS.	Negative
	@BarackObama : **Please** recognize the men and women who are occupying wall street.	Negative
	Obama suggests MLK Jr. **would have** backed #occupywallstreet.	Positive
#obamacar	**We need to** completely repeal #Obamacare and start by replacing it with HR 3400! #cnndebate	Negative
	If ObamaCare is not repealed **then** we can **expect** stagnant growth, long term unemployment and record high premiums.	Negative
	Op-ed: **Despite** conservatives' claims, #Obamacare is having little impact on hiring, writes Dean Baker http://t.co/I4pCBEOA #p2	Negative
#jobsnow	**Thanks to** @BarackObama's efforts, 270 businesses have committed over 25,000 jobs to American veterans.	Positive
	Cain creates 11 more jobs in 1 day **than** Obama in a lifetime.	Negative
	Let's be clear, the US economy is horrible **because of** Obama's policies.	Negative

Table 12: A case study of the opinion summary for "Obama"

[5] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] J. Cheung, G. Carenini, and R. Ng. Optimization-based content selection for opinion summarization. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 7–14, 2009.

[7] Y. Choi and C. Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proc. EMNLP*, pages 793–801, 2008.

[8] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proc. COLING*, pages 241–249, 2010.

[9] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972, 2007.

[10] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proc. ACL-HLT*, pages 42–47, 2011.

[11] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. KDD*, pages 168–177, 2004.

[12] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proc. AAAI*, 2004.

[13] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proc. ACL-HLT*, pages 151–160, 2011.

[14] H. Kim, k. Ganesan, P. Sondhi, and C. Zhai. Comprehensive review of opinion summarization.

[15] C. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)*, pages 25–26, 2004.

[16] B. Liu. Opinion mining. In *Encyclopedia of Database Systems*, pages 1986–1990. 2009.

[17] C. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. 2008.

[18] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proc. ACL*, pages 432–439, 2007.

[19] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proc. WWW*, pages 171–180, 2007.

[20] T. Nakagawa, K. Inui, and S. Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Proc. NAACL-HLT*, pages 786–794, 2010.

[21] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. ICWSM*, pages 122–129, 2010.

[22] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.

[23] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. EMNLP*, pages 248–256, 2009.

[24] M. Resende and R. Werneck. A hybrid heuristic for the p-median problem. *Journal of Heuristics*, 10(1):59–88, 2004.

[25] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proc. ACL-HLT*, pages 308–316, June 2008.

[26] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proc. WWW*, pages 111–120, 2008.

[27] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proc. CIKM*, 2011.

[28] L. Zhuang, F. Jing, and X. Zhu. Movie review mining and summarization. In *Proc. CIKM*, pages 43–50, 2006.