Summarization of Business-related Tweets: A Concept-based Approach

Annie LOUIS[1] *Todd NEWMAN*[2]
(1) University of Pennsylvania, Philadelphia PA 19104, USA
(2) FUSE Labs Microsoft Research, Redmond WA 98052, USA
lannie@seas.upenn.edu, todd.newman@microsoft.com

ABSTRACT
We present a method for summarizing the collection of tweets related to a business. Our procedure aggregates tweets into subtopic clusters which are then ranked and summarized by a few representative tweets from each cluster. Central to our approach is the ability to group diverse tweets into clusters. The broad clustering is induced by first learning a small set of business-related concepts automatically from free text and then subdividing the tweets into these concepts. Cluster ranking is performed using an importance score which combines topic coherence and sentiment value of the tweets. We also discuss alternative methods to summarize these tweets and evaluate the approaches using a small user study. Results show that the concept-based summaries are ranked favourably by the users.

KEYWORDS: tweets, twitter, summarization, business, concepts, domain-specific.

Proceedings of COLING 2012: Posters, pages 765–774,
COLING 2012, Mumbai, December 2012.

1 Introduction

In this work, we focus on tweets that mention a company name. Such company-related tweets are useful to multiple audiences. Tweets are a good source of public opinion. Hence company analysts and internal users can benefit from an overview of social chatter about the company. On the other hand, consumers are interested in reviews about a company for product, job-related and financial aspects. Other non-opinion content in tweets such as deals, job postings and advertisements are also useful to consumers. But the volume of tweets and their unconnected nature make browsing a stream of tweets rather difficult. This paper explores how to categorize tweets into subtopics and create a representative summary for each subtopic.

The challenge for this task is the diversity of the tweets. Tweets related to a company range from current news involving the company to job postings, advertisements, and cursory mentions. Moreover, tweets are short and contain informal language. As a result, there is little word overlap between tweets making it difficult to categorize them. We introduce an innovative method that performs broad clustering and does not rely solely on word overlap. Central to the method is the automatic acquisition and use of business-specific concepts. Our three-step approach is briefly summarized below:

1. Concept learning. Firstly, we acquire possible business concepts which are related to any company. For example, a company would have people in its management, customers, products, stocks and financial matters, events related to the company etc. Each of these 'people', 'products', 'assets' and 'events' could be a possible aspect for dividing the tweets. Our innovation is to learn such a set of business aspects automatically from an external source other than tweets—business news articles. Each concept is a group of related words identified from business news articles but also includes flexibility to handle new words in tweets that were unseen during concept extraction. Further this procedure is done offline only once and does not rely on any tweets.

2. Tweet clustering. All companies are assumed to have the same set of concepts indentified above. The tweets for each company get mapped to these concepts forming clusters. This mapping process allows even tweets with non-overlapping words to map to the same cluster.

3. Cluster ranking and summarization. These clusters are ranked using properties such as influential subtopic and sentiment associated with it. For this purpose, we also develop a sentiment classifier for business tweets.

We compare our method with other ways of summarizing the tweets and provide a small annotation study to understand user preferences. We found that the concept-based approach is able to provide useful summaries of tweets.

2 Dataset and types of business tweets

This section describes how we obtain the input tweets for our summarization system. We used an existing Microsoft crowdsourcing framework to obtain keywords related to different companies. We gave a company's name and asked people to add any keyword related to the company. Most keywords were related to products, people in its management, and affiliated businesses. There was a maximum of 5 keywords for a company and we also include the company name in keyword set. Each keyword is used to collect matching tweets from the past three days. The set of tweets for *all* the keywords for each company is the collection we wish to summarize for that company. The number of tweets for the companies in our development and test sets are shown in Table 1.

The tweets vary in their source as well as content. Some broad categories are shown in Table 2.

Development set						Test set	
Bank of America	695	RBS	197	Supervalu	108	Wells Fargo	1020
Sams Club	359	Costco	140	Abbot Labs	97	Lowes	887
JP Morgan Chase	351	Comcast	150	Sage Summit	87	Johnson & Johnson	811
Samsung	314	Delta Airlines	129	Att wireless	68	Northrop Grumman	280
Exxon Mobil	287	Prudential	128	Trader Joes	33	LinkedIn	280
Goldman Sachs	256	Safeway	125	Easy Jet	39	Nokia	158

Table 1: Companies and number of associated tweets

1. Related to news
RT @user1: Goldman Sachs: Calling for Greater Oil Price Speculation, Again http://.../ "vampire squid", indeed

2. Comments on products/services
Walmart orange chicken is digustin!!! My mom learned her lesson only the SAMs club version now on

3. On company aspects (eg. financial matters, people)
Sen. Rubio: we don't need new taxes, we need new taxpayers Agreed, how about we start with GE and Exxon/Mobil.

4. Comments not related to any particular aspect
I JUST LOVE WHEN BANK OF AMERICA LIESSSS TO MEEEE!

5. Postings from other applications such as 4Square
I'm at JPMorgan Chase in Lake Mary, FL http://

6. Advertisements, job postings
AZ Jobs | North Phoenix- Part time Teller - 67th Ave

7. Mentions but not really about the company
My little bro just asked me was uncle Sam the owner of Sams club...

Table 2: Types of business-related tweets. The company-related keyword is underlined.

3 Related work

Most methods for summarizing tweets have either focused on tweets matching a generic search query (O'Connor et al., 2010), or on tweets related to sports and celebrity events (Sharifi et al., 2010; Chakrabarti and Punera, 2011; Liu et al., 2011; Inouye and Kalita, 2011; Nichols et al., 2012). We focus on summarizing the tweets that show up in the search for a company.

In fact, our work is more related to aspect-based summarization methods commonly employed on product reviews (Hu and Liu, 2004; Gamon et al., 2005; Sauper et al., 2011; Zhai et al., 2011). These methods first obtain a set of attributes for the product. For example, for a camera, the attributes may be "lens", "focus" and "zoom". Then positive and negative sentences in the reviews are divided according to these attributes and their aggregate are shown for each category. Some approaches obtain the attributes through manual annotations and domain resources (Gamon et al., 2005; Zhuang et al., 2006). Others learn the attributes automatically using the review text (Hu and Liu, 2004; Titov and McDonald, 2008; Sauper et al., 2011; Zhai et al., 2011). Frequently occurring phrases in the reviews and also which often tend to be associated with sentiment as chosen as attributes. But while product review archives have significant overlap in topics, the informal nature of twitter conversation creates diverse tweets and also mixes review and non-review content. Identifying frequent attributes from tweet streams becomes difficult and unreliable. So we use an external resource, news articles, to learn concepts for the business domain and use these concepts to guide clustering of tweets. Our procedure for learning concepts is fully automatic and without reference to individual companies whereas product review attributes are usually specific to the product.

Our cluster ranking procedure is also novel compared to prior approaches that either do not rank the clusters explicitly or use only sentiment information for ranking (Gamon et al., 2005; Zhuang et al., 2006; Blair-Goldensohn et al., 2008; Sauper et al., 2011). In our work, we merge sentiment information with a score to identify if a subtopic discusses an overwhelming issue.

packages	pay	bonus	profits	examples	wiring	kiosks
talent	telecommunication	stability	bookstores	cancellation	pilot	supplies
prize	investigation	equals	applicant	shutdown	savings	earnings
actuary	reports	plant	notice	vehicle	pilots	brands

Table 3: Samples from the company-word dictionary

4 Concept-based summarization

We present our three-step approach in this section.

4.1 Concept creation

This step creates a dictionary of business-related concepts using one year's worth of news articles from the New York Times (NYT) corpus (Sandhaus, 2008).

We first identify company names in these articles. A named entity tagger is used to automatically mark all mentions of 'organizations'. Using the metadata in the NYT corpus, we identify articles that appeared in the business section of the newspaper and only the 'organization' mentions in these articles are considered as possible company names. These company names are replaced with a generic token "COMPANY" because we are interested in words associated with company mentions in general without reference to individual companies.

Then the nouns (proper nouns are excluded) in a window of 20 words each before and after all COMPANY tokens are obtained as a list of candidates for the dictionary. For each candidate word w_i, we compute its association with COMPANY tokens in the corpus using mutual information.

$$\text{MI}(w_i, \text{COMPANY}) = \log \frac{p(w_i, \text{COMPANY})}{p(w_i)p(\text{COMPANY})}$$

p(w_i, COMPANY) is the probability with which w_i is found in the vicinity (20 word window before and after) of COMPANY tokens. p(w_i) is the probability of w_i in the full corpus and p(COMPANY) is computed likewise.

The top 2000 nouns in this ranking are selected to create a company-word dictionary (a random sample is shown in Table 3). Next we group these words using WordNet (Miller, 1995) to obtain more general concepts. We obtain the list of synsets on the hypernym path between each company-word and the root of WordNet. Then we record the synset names for a word at levels 3, 4 and 5 from the root. (Root is considered as level 1.) The sequence of these 3 synsets is considered as the SEMANTIC TAG for the word. Word that map to the same SEMANTIC TAG are grouped and correspond to a concept. We choose levels 3, 4 and 5 to obtain a concept that is neither too specific nor too general. The resulting set has 57 diverse concepts and most of them can be intuitively understood to be business-related. We manually assigned a name for each concept based on the SEMANTIC TAG and the group of words. Each concept is a triple (T, L, D) where T is its SEMANTIC TAG and L is the MANUAL LABEL. D represents the grouped words (called PRIOR WORDS) for that concept. Table 4 shows example concepts with different number of PRIOR WORDS.

All the above processing is done offline and only once. Note that up to this step, we have used only the news articles and WordNet for concept extraction.

4.2 Mapping tweets into concepts

For each company, we assume that the same set of 57 concepts are the possible subtopics for its tweets. We assign each tweet to one of these concepts.

Semantic tag [T] (Levels 3 - 4 - 5)	Size	Example prior words [D]	Manual label [L]
[psychological feature] - [event] - [human activity]	341	merger, consultancy, takeover	Activities
[physical object] - [unit] - [animate thing]	208	acquirer, creditor, sibling, analyst	People
[physical object] - [unit] - [artefact]	189	airline, appliance apparel, auto	Artefacts
[group] - [social_group] - [organization]	54	carmaker, insurer, division, firm	Group
[matter] - [substance] - food	23	beer, provisions, candy, snack	Food
[relation] - [possession] - [property]	5	trust, effects, estate, property	Property
[attribute] - [quality] - [asset]	5	specialty, asset, advantage	Plus/Quality

Table 4: Some company-related semantic concepts. 'Size' indicates total number of prior words.

This process involves computing a membership score for each tweet and concept pair (t_i, C_k). The score has two components—exact and fuzzy. We first record words from t_i which directly match any of the PRIOR WORDS D of the concept C_k. We call these words as exact matches, set E, for that concept. For each of the remaining words in the tweet, we compute its SEMANTIC TAG from WordNet as before and check if it matches the tag T of C_k. In the event of a match, we add the word to the set of fuzzy matches F. The remaining words are ignored. The membership score for the tweet-concept pair is computed as:

$$score(t_i, C_j) = \lambda * |E| + (1 - \lambda) * |F|$$

Here λ is set to 0.8 to give higher weight to exact matches. The union of exact and fuzzy matches $E \cup F$ are stored as the MATCHING WORDS for that tweet-concept pair. The tweet is assigned to the concept with which it has maximum membership score. Where there is a tie, the tweet is assigned to the concept that has the most non-zero membership values across all tweets. In this way, the tweet is assigned to the more general of the candidate concepts.

4.3 Cluster ranking and summarization

We summarize the resulting clusters using two modules.

4.3.1 Cluster ranking

We introduce a method to rank clusters by combining sentiment value and entropy of word distribution in the cluster. The intuition is that when the tweets in a cluster discuss a common issue, we should rank it higher than a cluster which has diverse content. For example, on a given day when the CEO of a company resigns, many users discuss the event and so the "people" concept cluster of the company would have homogenous content on that day. In addition, the tweets in such a cluster will also have a lot of sentiment.

We use the entropy of the word distribution in a cluster as a measure of homogeneity and also adapt the score to consider the sentiment of words. Further, rather than use all the words in the cluster, we utilize only a smaller set of topical words which we obtain by combining all the MATCHING WORDS (see Section 4.2) for tweets belonging to that cluster.

Consider a cluster C_j and the union of MATCHING WORDS for its constituent tweets is the set M. The probability of a word $w_i \in M$ is given as:

$$p(w_i) = \frac{wtcount(w_i)}{\sum_{w_k \in M} wtcount(w_k)}$$

where $wtcount(w_i) = \sum_m sentimentValue(S_m)$. Here S_m is a tweet MATCHED to C_j by w_i.

The sentiment value of a tweet ranges between 0 and 1 and is obtained from a sentiment classifier. The classifier does a 3-way division of tweets into positive, negative and neutral

(+57) People: customer, banker, employee, man
Wells Fargo holding a daylong seminar to help customers having problems with mortgage http://../
I swear Wachovia care more about customer service than anything
Wells Fargo decided to exit reverse mortgages after federal officials insisted it foreclose on elderly customers http://../
(+14) Amount: money, cash, fund
Wells Fargo act like they are mad about their little money
Wells Fargo lost cash off my card. #smh I am sewing someone

Table 5: Snippet from a concept-summary for *Wells Fargo*. (+x) indicates cluster size.

categories and outputs a probability distribution over these 3 classes. The sentiment value is the absolute difference in positive and negative confidence value from the classifier. This score indicates the degree to which the tweet is oriented towards one kind of sentiment—positive or negative and takes the highest value of 1 when the tweet is predicted as fully positive or negative. Using these sentiment-aware probabilities, we compute the entropy of C_j.

$$H(C_j) = -\sum_i p(w_i) \log p(w_i)$$

Lower values of entropy indicate a skewed distribution of MATCHING WORDS and therefore a better cluster. But a large cluster is likely to get higher entropy even if it is cohesive, compared with a smaller cluster. So we apply a weighting factor to reduce the entropy of large clusters.

$$H_{adjusted}(C_j) = (1 - \frac{|C_j|}{\sum_k |C_k|})H(C_j)$$

This score $H_{adjusted}$ is the final score for a cluster. Lower scores indicate higher ranked clusters.[1]

4.3.2 Faceted summarization

This step generates a summary for the top-ranked clusters. First we obtain the top four MATCHING WORDS of the cluster that have highest probability (also incorporating sentiment as in the previous step). These words are displayed as a headline for the cluster.

For each headline word, we identify all the tweets containing that word. We compute average probability of words in each tweet and rank them in descending order of score. The average probability scoring is a popular and successful method for automatic summarization (Nenkova et al., 2006). The probability value is computed as in the previous section by also incorporating sentiment information. We only use the first two headline words for summary generation. For the first headline word we pick the top two sentences from its ranked list and we choose one sentence for the second word. For the final interface, the clusters are shown in rank order up to a certain limit on the number of tweets displayed. Table 5 shows an example summary.

5 Sentiment classification

We built a 3-way sentiment classifier for our task. We annotated 2470 tweets from the development set as positive, negative or neutral in sentiment. Exact retweets were removed and when the main topic of a tweet was not the company, it was annotated as neutral regardless of other sentiment. Annotators include the authors and six software engineers. The resulting data had 49.5% neutral, 22.8% positive and 27.6% negative tweets.

[1]But when the entropy is zero (only one MATCHING WORD), we lose information about sentiment value. For such zero-entropy clusters, check the average sentiment value on the MATCHING WORD and if below a threshold, we demote the cluster and assign to it the largest entropy value across all clusters.

CN is first word of tweet	CN has dependency link to main/some verb
POS of words within two words around CN	CN has positive/negative modifier or sibling
POS of CN's parent	Modifier of CN's sibling is positive/negative
Sentiment of CN's parent and grandparent	Positive/Negative word within two words around CN

Table 6: Target-based features for sentiment prediction

Our features include counts of unigrams, bigrams as well as parts of speech (POS) tags and punctuations. We also count the sentiment words using two lexicons (MPQA (Wilson et al., 2005) and General Inquirer (Stone et al., 1966)) and a hand-built dictionary of sentiment-related slang words. We also added features specifically aimed to identify if the company is the main target of the tweet. These features were computed from a dependency parse of the tweet and are briefly listed in Table 6. In this list, 'CN' indicates the company keyword present in the tweet. We used a MaxEnt classifier for training and performed 10-fold cross validation.

The n-gram, sentiment words and POS features gave an accuracy of 64%. Target-based features increased the accuracy to 82.6% showing that such features are valuable for our task.

6 Alternative summarization methods

We introduce three other methods of summarizing tweets for comparison with our approach.

a) Sentiment only (Sen). A simple summary for our task is showing the top positive and negative tweets (according to classifier confidence).

b) Frequency only (Frq). This summary aims to show the most discussed tweets in the stream. For each tweet, we compute the number of similar tweets. Two tweets are considered similar when the cosine similarity based on unigram counts is above 0.8. The tweets with the largest number of similar tweets are displayed along with the number of similar tweets.

c) No categorization but sentiment + frequency (Prb). We apply the same summarization method as used in our concept method. The probabilities of words (also using sentiment) are computed over the full set of tweets. Then sentences are ranked by the average probability of words. But the sentences are not categorized into positive/negative or frequency sets. The average probability method works remarkably well for summarizing newswire (Nenkova et al., 2006), the domain where more mature systems exist. So we include it for comparison.

Table 7 shows snippets from alternative summaries for "Wells Fargo".

7 Annotation experiment

For each of the four approaches, we generated summaries containing a maximum of 20 tweets. In the case of the concept approach (**con**), this limit is for the total tweets across all clusters.

1. Sentiment summary (Sen)	**2. Frequency summary (Frq)**
(+) I love Wells Fargo. They let you customize your debit card!	(+19) Banks financing Mexico drug gangs admitted in Wells Fargo deal
(-) Wells Fargo has pissed me off one too many times. Time to move my money	(+14) Wells Fargo to pay $125 million in mortgage suit http://t.co/

3. Average probability summary (Prb)
Wachovia banks to become Wells Fargo
Wells Fargo, Goldmann Sachs and all other banks don't come close

Table 7: Snippets from other summary approaches for "Wells Fargo". (+) and (-) indicate polarity. (+x) indicates cluster size.

	Useful for analysts			Informative for consumers			Interesting for consumers		
	sen/con	**prb/con**	**frq/con**	**sen/con**	**prb/con**	**frq/con**	**sen/con**	**prb/con**	**frq/con**
well fargo	sen	**con**	**frq/con**	sen	**con**	frq	sen	**con**	frq
johnson	**con**	**con**	**con**	**con**	**con**	-	sen	**con**	-
linkedin	sen	**con**	**con**	sen	**con**	-	sen	**con**	frq
nokia	**con**	**con**	frq	**con**	**con**	frq	**con**	-	frq
northup	sen	**con**	-	sen	**con**	**con**	**con**	**con**	**con**
lowes	**con**	prb	**con**	**con**	prb	**con**	**con**	prb	**con**
% con		61.1			55.6			50.0	

Table 8: Evaluation results. The header indicates the pair that was compared and cells indicate user judgement. '-' denotes no preference and x/y indicates both x and y are preferred.

We use the 6 companies listed in Table 1 as the test set. For each company, we paired the output of the concept approach with each of the alternative summaries. Judges were asked to provide their preference between the summaries in each pair. Our judges were 14 software developers and had no prior computational linguistics experience. Each judged two or three random pairs of summaries and did not see more than one pair from the same company. They were asked to answer three questions.

If you were an analyst working for the company,
Q1) Which summary would be more useful for you?
Imagine you are a consumer interested in learning about a company. From your viewpoint,
Q2) Which summary was more informative? It gave you a useful overview about the relevant tweets.
Q3) Which summary was more interesting to read?

The judges had 4 options "summary A", "summary B", "prefer both", "none". Table 8 shows the judgements provided for our test set. The last row indicates for each question, how often the concept approach summary was preferred in the 18 judgements that were made.

In the analyst view, concept summaries are highly preferred. 61% of the comparisons noted this summary as better than an alternative method. For informativeness quality, the concept summary was preferred 55% of the time and 50% of the cases for interest value. When all three questions are put together, there are 54 judgements and the concept summary was preferred 30 times, 55%. Our test set is small, still these results indicate that judges find the concept summaries useful. The concept summary was almost always better than the PRB option where there was no clustering into subtopics. But judges noted that the SEN summary was fairly intuitive and easy to interpret.

8 Conclusion

We showed that use of domain concepts can provide a useful summarization method for diverse tweets. Since we only rely on unannotated news articles and WordNet which are available in other languages as well, our method is also easily portable. Another attractive feature of our approach is that the same concepts are used for all companies. So one could track what happened in the "people" cluster across different companies or over time for the same company. On the other hand, fine-grained concepts for different classes of companies such as technology versus finance could also be interesting to obtain. We plan to explore these ideas in future.

We also found that properties of the tweet stream influenced the quality of the summary. Some companies' tweets were mostly offers and deals and here concept summaries were less useful. Frequency or sentiment summaries displayed more interesting tweets. So we want to explore how to vary the summarization approach depending on the type of tweets in the input set.

References

Blair-Goldensohn, S., Neylon, T., Hannan, K., Reis, G. A., Mcdonald, R., and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In *In NLP in the Information Explosion Era*.

Chakrabarti, D. and Punera, K. (2011). Event summarization using tweets. In *Proceedings of ICWSM*.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. K. (2005). Pulse: Mining customer opinions from free text. In *Proceedings of IDA*, pages 121–132.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of AAAI*, pages 755–760.

Inouye, D. and Kalita, J. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *Proceedings of IEEE Third International Conference on Social Computing*, pages 298 –306.

Liu, F., Liu, Y., and Weng, F. (2011). Why is "sxsw" trending? exploring multiple text sources for twitter topic summarization. In *Proceedings of the ACL Workshop on Language in Social Media (LSM 2011)*, pages 66–75.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communication of the ACM*, 38(11):39–41.

Nenkova, A., Vanderwende, L., and McKeown, K. (2006). A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of SIGIR*.

Nichols, J., Mahmud, J., and Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the ACM International Conference on Intelligent User Interfaces*, pages 189–198.

O'Connor, B., Krieger, M., , and Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of ICWSM*.

Popescu, A. and Jain, A. (2011). Understanding the functions of business accounts on twitter. In *Proceedings of WWW*, pages 107–108.

Sandhaus, E. (2008). The new york times annotated corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia*.

Sauper, C., Haghighi, A., and Barzilay, R. (2011). Content models with attitude. In *Proceedings of ACL-HLT*, pages 350–358.

Sharifi, B., Hutton, M., and Kalita, J. (2010). Summarizing microblogs automatically. In *Proceedings of HLT-NAACL*, pages 685–688.

Stone, P., Kirsh, J., and Associates, C. C. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Titov, I. and McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL*, pages 308–316.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354.

Zhai, Z., Liu, B., Xu, H., and Jia, P. (2011). Clustering product features for opinion mining. In *Proceedings of WSDM*, pages 347–354.

Zhuang, L., Jing, F., and Zhu, X. (2006). Movie review mining and summarization. In *Proceedings of CIKM*, pages 43–50.