Social Topic Analyzer

Aastha Nigam
University of Notre Dame
anigam@nd.edu

Salvador Aguinaga
University of Notre Dame
saguinag@nd.edu

ABSTRACT

Media communications companies interact with their follower base on social media and other mediums. Online social media platforms provide a feedback mechanism for content providers to know how specific content is being shared across follower groups. For media companies to keep followers engaged and potentially attract new followers to their digital content, they are interested in what their followers are talking about. In this work, we aim to understand what followers on social media are talking about, but media companies are not covering. By probing the social media network for emerging terms and topics, we also aim to leverage prolific tweeters' social network to identify and potentially attract new followers while keeping current ones engaged. Moreover, we present a new framework that addresses these aims and evaluate it on Twitter followers of Schurz Communication - a small media communications company.

1. INTRODUCTION

Twitter is a micro-blogging site which helps user in exchanging information. Many companies or prominent users use this platform to connect with others. Users can follow other users they are interested in interacting with or consuming information from. Many newspapers have their Twitter pages which they use to share their articles. The influence is captured by the number of users following that newspaper or re-tweeting their content or commenting on it. Twitter captures the trending topics over a period of time. In a similar setting, it will be extremely helpful for a newspaper to know what is trending among its users so that they can provide information that would keep their customers engaged. Since, a tweet is restricted to 140 characters understanding a category such as Food and Drink or Art and Music becomes difficult. Therefore, we can model the topic from the tweets using the text provided.

In this paper, we work with data obtained for a local news media company, Schurz Communication (SCI) Inc. [1]. This company consists of seven business segments: Broadcasting Radio, Broadcasting TV, Cable TV, Newspaper Publishing, Shoppers, and Digital Media. Two of their most popular digital properties are the South Bend Tribune (SBT) newspaper and a television station (WSBT-TV). One of the technologies SCI uses to connect with their customers is the social media platform: Twitter. Both properties currently have a sizable follower base, 12.1 and 20 thousand followers respectively. News reporters for the two businesses tweet information based on their understanding of what is important and trending. One of the problems addressed in this work is: what are their followers talking about, but is not being covered by SCI. There are multiple sources of news on Twitter such as New York Times, Washington Post and Chicago Tribune for a newspaper reader. To understand what the user will be most interested in reading about we look at the users' tweeting patterns. Secondly, another problem we study is the identification of potential new followers. This problem stems from SCI's interest to increase their follower base by attracting followers that are most likely to be interested in reading or consuming media content offered by them. In this work we study how to leverage the network of random and prolific tweeters following WSBT and SBT to identify the overlap of topics or terms that emerge in this sub-network of Twitter users. Prolific tweeters are those followers identified by the media company as individuals that retweet content, tweet about content, and have a network of friends and followers on Twitter that is significant when compared to a typical or an average follower. Therefore, we propose a topic-based system with the following two objectives:

1. **Topic modeling on tweets by SCI followers**: We aim to understand what the followers of SCI are talking about on Twitter. This when contrasted with the information SCI provides on its Twitter pages will help them understand what their followers are most interested in reading about. This in turn will lead to increased engagement of their followers. To achieve this we propose to do topic modeling on the tweets obtained from their current followers. Using this method we will be able to build a topic based profile for an average user on Twitter for SCI.

2. **Recommendation of potential followers**: We aim to find potential followers for SCI. We aim to leverage the friends' network of the current prolific followers as a candidate set for potential followers. Using our topic based user profile, we aim to find users which have the best overlap with the content currently provided by SCI. We claim that users who are reading/tweeting about topics that SCI is sharing and are not current SCI followers are the potential followers for SCI.

Thus, our objective is to design and implement a system that

[1] http://www.schurz.com/

identifies what their followers are talking about and determine if these topics are not being covered by SCI. Secondly, we plan to utilize the outcomes of the system to develop a model to identify potential new followers, from current SCI followers' friend network, who might be interested in SCI content.

The paper is divided into the following sections: We firstly describe other research that has been done on topic modeling on Twitter in Section 2. Then in Section 3, we describe our data collection process from Twitter and Section 4 briefly talks about data pre-processing and the cleaning tasks. We then explain our entire framework and propose a model for topic identification in tweets in Section 5. Section 6 describes our two experiments for user based topic profiling and recommendation of potential followers. We also present a mobile application to be used in real time as described in Section 7. Lastly, we conclude in Section 8.

2. RELATED WORK

One of the first approaches used to understand the topic in a tweet was based on the term frequency using tf-idf [6]. One of the most popular and useful algorithms for topic modeling is LDA [4]. But it is unable to perform well in the context of tweet because they contain only 140 characters [5]. These two methods try to learn the topic from the content of the tweet. Another approach that is useful is to use external sources to understand the topic of the tweet. Bernstein et. al [2] proposed a technique called TweeTopic that uses search engine as a distributed knowledge base. They use parts-of-speech tagger on the tweet to create a query which is fed into the Yahoo! Build Your Own Search Service. The pages returned from this search are used to understand the topic. The popular words in the search result are calculated using inverse frequency document and are used as the topics. Similarly, Macskassy et. al [1] use the entities in a tweet for understanding the higher level categories. They retrieve Wikipedia [2] articles for each query and create a tree using the categories. They identify the category of an entity using this tree and its context using the tweet.

3. DATA COLLECTION

The collection of data was sourced from Twitter, where a select number of SCI's digital properties with Twitter accounts had their timelines mined. The data was collected to serve as the *base-dataset* which reveals what the digital property is tweeting about. A second component of the dataset was to collect the set of tweets ($\mathcal{X}$) from their followers and divide them into two categories: tweets from prolific tweeters and from typical tweeters ($p, t \in \mathcal{X}$). The aims were to identify what highly engaged Twitter users microblog about and learn how large the intersection of topics is between the prolific tweeters and SCI's digital properties. In connection to their most prolific tweeters, another goal was to explore potential new followers and consumers of SCI's content. SCI collected tweets for periods of time by focusing on their prolific tweeters and shared these data for the purpose of this study.

In terms of content consumption, typical tweeters have potential to not only become more engaged with SCI's content, but to potentially influence their social network as well. For this group, we selected SCI's followers at random to obtain tweets from average followers. The collection of these data was performed by selecting a set of average followers at random, $f_{rnd} \in \mathcal{X}$, and fetching their most recent ten tweets. This process was repeated by selecting new sets of followers every hour for a few days.

Base-Dataset Tweets	
Property	Count
SBT	1425
WSBT	1335
Prolific Tweeters	
SBT	1878
WSBT	1885

Table 1: Number of tweets collected for this dataset.

4. DATA PREPROCESSING

As described in Section 3, the tweets were obtained for random and prolific users. The data retrieved from Twitter was in JSON format. We first parsed it into a comma separated file keeping only the required fields such as tweet ID, time it was created, the text from the tweet, the number of times it has been retweeted, the urls shared in the tweet, the hash-tags, the users mentioned in the tweet, favorite count and if the tweet is in reply to some other user. Once we had these fields, we mostly focused on the tweet ID, the text and the urls. A tweet typically has many unnecessary characters and expressions which do not contribute much to a tweet. Therefore, we cleaned a tweet of its unnecessary characters using regular expressions. We understand that emoticons or expression marks can say a lot about the sentiment of the tweet, but since the focus of this research is mainly to understand the topic category for a tweet, emoticons and other punctuation marks are not important for this study. Figure 1 shows a brief example of different stages of preprocessing.

5. MODEL DEVELOPMENT

As described in the previous section, we collected tweets from SBT and WBST Twitter followers. The text of the tweet was extracted. Since our goal is to understand the topics captured in a tweet we firstly discover the entities in a tweet. We begin by discovering the named entities in a tweet. This process is called Named Entity Recognition (NER) [8]. We focus on the Parts-Of-Speech (POS) tagger which tags every word in a sentence to a parts of speech based on it's definition and context. There are many popularly available POS tagging algorithms but they suffer with tweets. The 140-character limit in a tweet does not provide sufficient information (context) to identify an entity. We initially used the Stanford POS Tagger [10] which has proved to produce good results but failed to correctly identify on our dataset. Therefore, we use CMU tweet tagger [9] which has been trained over a set of tweets. Once each word in a tweet was tagged by a parts of speech, the words marked as nouns were picked up from the tweet. We assume that the nouns

[2] https://www.wikipedia.org

Tweet:
’I have completed the quest ’Another Try’ in the #Android game The Tribez. http://t.co/uHpflTNi57 #androidgames, #gameinsight’
Prepressed an cleaned:
i have completed the quest ’another try’ in the android game the tribez. URL androidgames, gameinsight
Broken down to its components:

```
[(u’i’, ’PRP’),(u’have’, ’VBP’),(u’completed’,
’VBN’),(u’the’, ’DT’), (u’quest’, ’JJS’),
(u"’another", ’JJ’), (u’try’, ’NN’),
(u"’", "’’"), (u’in’, ’IN’), (u’the’, ’DT’),
(u’android’, ’JJ’), (u’game’, ’NN’), (u’the’,
’DT’),(u’tribez’, ’NN’), (u’.’, ’.’), (u’URL’,
’NNP’),(u’androidgames’, ’VBZ’), (u’,’, ’,’),
(u’gameinsight’,’NN’)]
```

Figure 1: Breakdown of a tweet from it raw text form to its tagged componets.

could represent the main context of the tweet. Using the above mentioned example, we see the words ’game’,’tribez’ and ’gameinsight’ are marked as nouns.

As a user, the most common step we do when we read a word which we are not familiar with is to query the word in a search engine. We replicated the same action. We built a query using the nouns extracted from each tweets. Since the order of the words in a query matter, we use the same order in which they occurred in the tweet. This query is fed into the Google search engine. For some tweets that do not have any nouns, we query the entire tweet text. For each query, we retrieve the top 20 links that Google provides. We do lose the context of the nouns by eliminating other words but we argue that if a particular word occurs in a tweet the most common or popular context will be captured by the top 20 results from the search engine. By this step, we aim to enrich the tweet text, which itself does not mean much, with publically available information on Google.

After retrieving top 20 results (or lesser) for each tweet, we crawl each url to get the content on that page. Our motivation behind this is to enrich the content of the tweet using external sources since 140 characters are not enough. We then clean the content returned by the urls of any unnecessary characters using regular expressions. Secondly, we remove the stop words such as ’a’, ’and’ and ’the’. Also, we remove all those words that just occur once in our data. For each tweet, we therefore have data from 20 urls which is condensed into a single document. In addition, we perform feature selection using a TF-IDF approach. We measure the tf-idf score for each word. We keep a very high threshold of tf-idf score and all words with value less than that are removed from the document. We tried various threshold values and by experimentation found out that at a 95 percentile is a good threshold. The same approach was repeated for each tweet and its corresponding 20 documents. Using these documents we build a dictionary of words in our dataset and represent each tweet in a document vector format. We then perform Latent Dirichlet Allocation (LDA) [4], a popular topic modeling approach, on the document corpus. LDA represents each document as a mixture of topics. Also, it gives a probability distribution of words constituting a topic. As a result, we get a list of topics over our entire tweet dataset and also get the topic contribution in each tweet.

Figure 2 captures our entire framework. Our framework consists of the following broad steps:

1. Clean the tweets of any unwanted characters and extract the text.
2. Identify the nouns in a tweet using a POS tagger.
3. Generate a query using the nouns in the tweet (keeping the same implicit order).
4. Feed the query into any search engine such as Google and obtain the top 20 results(url).
5. Crawl each of the urls returned for each tweet to create a document.
6. Perform basic preprocessing on the document by removing stop words.
7. Perform feature selection using tf-idf score.
8. Use the processed document as the corpus to train LDA.
9. Identify the topics in each tweet based on the results of LDA.

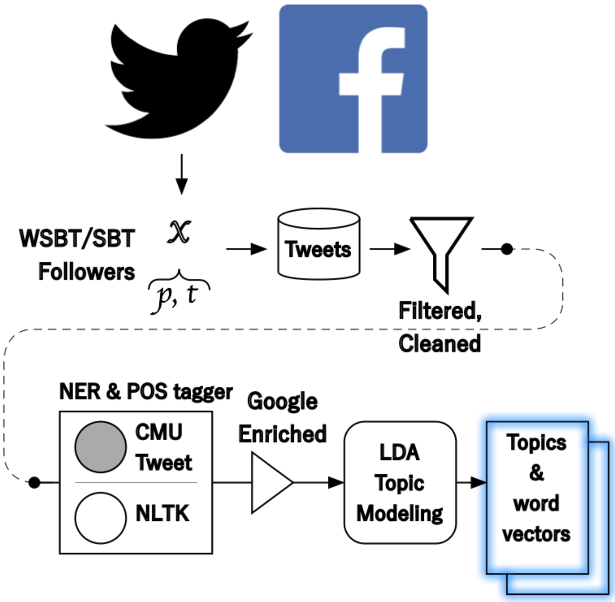

Figure 2: Framework: An overview of our framework showing the implementation system design for Twitter followers. $\mathcal{X}$ is the set of tweets that comprise our dataset, where some tweets come from prolific (p) and typical (t) followers.

6. EXPERIMENTS

This section describes the two experiments we conducted using the framework described in the previous section.

6.1 Average user topic profile

As described in Section 3, we collected tweets from random users and for the most prolific users of SBT and WSBT-TV. For this experiment, we will be looking at an average user rather than a prolific user. Since the aim of this experiment is to understand what the followers are talking about we want it to represent an average user. For proof of concept we work with 880 tweets collected from both SBT and WSBT-TV followers. We follow the framework described in Section 5. We cleaned the 880 tweets, identified the nouns and used them to make a query. We used the Google search engine to retrieve the top 20 urls. We obtained the text from these 20 urls and condensed them into one document by per tweet. We used this to create our corpus. LDA was trained on this model. We have set the number of topics as 10 for our experiments. We chose 10 topics since it is easily understandable and readable by users and can be easily displayed on visualization application (described later in Section 7). Table 2 lists only two topics out of the 10 we have. It also lists the words that constitute the topic and their probabilities. It can be seen that the words in topic 1, essentially are related to national politics whereas the second topic is related to the Ferguson case[3].

Table 2: Captures two topics obtained over the dataset by running LDA. With in each topic, we show the top words with their probabilities

National Politics		**Ferguson**	
word	**probability**	**word**	**probability**
immigrants	0.05	police	0.086
obama	0.004	camera	0.055
immigration	0.041	juries	0.003
president	0.003	brown	0.023
nation	0.018	officers	0.013
country	0.016	body	0.043
workers	0.021	wilson	0.021
illegal	0.002	prayer	0.002
undocumented	0.03	riot	0.22
employees	0.06	ferguson	0.11
security	0.07		

We also obtain the topic distribution for each tweet. We use this to build a user profile. We list all the tweets by the user and obtain the topics of those tweets. This helps us in getting an overview of all the SCI followers to understand what they are talking about and also us to have a more user-centric granular analysis.

We do a similar analysis on SBT and WSBT-TV. We crawled their timelines and obtained their tweets over the same time frame. Since, SBT and WSBT-TV are providing information by sharing their news articles and news videos, we assume they embed a url in their tweet. Therefore, we change our approach slightly here. Since we already have a list of urls, we do not make a query or search Google. We use their urls and get the content from their pages to build our document corpus. Once we have the corpus we apply LDA again to obtain the topics. This is done to understand the topics pertaining to the information provided by SCI. Once we have this it helps us to see the percentage of overlap between what their users talk about and what SCI provides.

6.2 Potential users

In this experiment, we try to find potential followers and recommend them to SCI. For this, we use the prolific users. Since, our aim is to increase the influence of SCI on Twitter we assume that it will be most effective if we consider the users who are the most active. To achieve this, we start by filtering the friend network of a follower to obtain a suitable candidate set. We obtain the friends of each current follower and filter based on whether or not they follow WSBT and SBT. Next, we look at their location. Since SBT and WSBT-TV mostly focus on Michiana area, we look for users in the given area. Thirdly, we look at only the prolific friends of the prolific users so that we can maximize our influence.

Once we obtain a suitable candidate set of users, we crawl their timelines to get their tweets. We then use our model described in Section 5 to build a topic based user profile. Based on the previous experiment, we also find the topic SCI talks about. We propose to use a simple majority vote to see which users talk about the topics SCI currently provides. If a user is already talking about topic x which SCI provides, he/she might be interested in subscribing to SCI too.

7. VISUALIZATION

This work explores an application-specific data visualization interface for mobile computing. Section 5 describes our data processing approach that yields various outputs organized for consumption on mobile devices, i.e. mobile phones and tablets. The organization of the output data centers around the presentation of *what followers are talking about.* To achieve this we present the fraction of topics that SCI's followers are talking about, but SCI is not. These topics are listed in both, a table-view and a word (or tag) cloud. When a user selects one of the trending topics the interface lists the actual tweets on the topic. The interface offers two other features: selection of trending topics by timeline and by user or prolific tweeter. There is a difference in presentation between trending topics and prolific tweeter visualizations. Selecting prolific tweeter views result in that user's tweets to be listed in the tableview and the topics within his', her's, or its' (for tweet-bots) tweets posts.

The visualization is prototyped on the iOS platform. Figure 3 shows screenshots of the actual implementation running on the iOS Simulator. The back-end for this app allows for real-time processing of input data. The framework and the topics model processes input and generates output in stream-like mode allowing for event driven data synchronization with clients (iOS devices running TalkBender). An other feature of the app, currently under exploration, is geo-location visualization showing tweet provenance. The mapview offers a picture of where tweets originate from. This type of information can be used to explore topics emerging around local news or events.

8. CONCLUSION

This work explores challenges in topic recommendation from online social media platforms such as Twitter. This work aims to identify what SCI's followers are talking about, but SCI is not covering on their media outlets. A second aim of this work is identify potential followers for SCI. To address the first problem we explore topic modeling to identify the

[3] http://en.wikipedia.org/wiki/2014$_{F}erguson_{u}nrest$

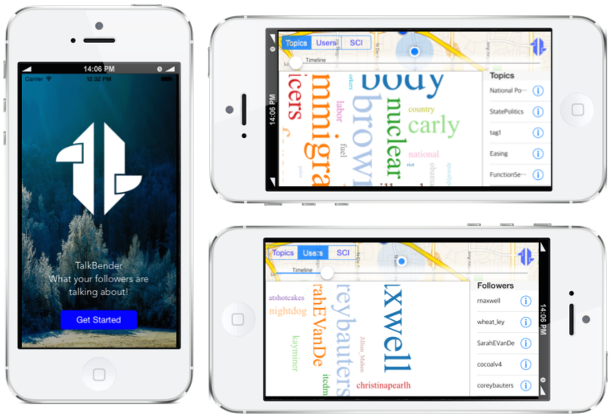

Figure 3: An iOS app-mockup plan for TalkBender. The app will show trending topics of what SBT/WSBT followers are talking about and will group the information by their most prolific followers and what their average followers are talking about.

terms emerging out of the tweets of those following WSBT and SBT. The latter problem is addressed by profiling followers of those in SCI's most prolific tweeter network to look for similarities in the topics they talked about with those emerging out of the most prolific follower group.

Our approach for topic modeling consisted of enriching tweets using external sources of public information in order to run LDA topic modeling on the resulting corpus. In this work we took a novel approach to visualization. We developed a mobile application to visualize the results of our model. The mobile application, TalkBender (see Figure **??**), runs native on iOS devices and offers the user a simple method to see emerging topics for for WSBT or SBT, as well as subtopics under a more general topic or term, and topics emerging by the users.

Future improvement to this work centers around social network sensing of emerging terms or topics across online social media platforms including Facebook and Google+. Another area for potential research work is on automatic connection or the mapping of emerging terms to concepts.

9. ACKNOWLEDGEMENTS

We would like to thank Schurz Communications for providing access to data.

10. REFERENCES

[1] Michelson, M. and Macskassy, S.A.: Discovering users' topics of interest on Twitter : A first look. In: Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (AND). Toronto, Canada (2010).

[2] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam and Ed H. Chi, Eddi: interactive topic-based browsing of social status streams, Proceedings of the 23nd annual ACM symposium on User interface software and technology, October 03-06, 2010, New York, New York, USA.

[3] Zhao W. X., Jiang J., He J., Song Y., Achananuparp P., Lim E-P. and Li X., Topical keyphrase extraction from Twitter, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 19-24, 2011, Portland, Oregon.

[4] Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 4-5 (2003), 993-1022.

[5] Ramage, D., Dumais, S., and Liebling, D. Characterizing Microblogs with Topic Models. ICWSM '10, AAAI Press (2010).

[6] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 5 (1988), 513-523

[7] Ramage D., Hall D., Nallapati R., and D. Manning C. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 248-256.

[8] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 1524-1534.

[9] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2 (HLT '11), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 42-47.

[10] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13 (EMNLP '00), Vol. 13. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI=10.3115/1117794.1117802 http://dx.doi.org/10.3115/1117794.1117802