TOPIC MODELS TO UNDERSTAND USER ENGAGEMENT ON TWITTER

Aastha Nigam[1,2,3,4], Salvador Aguinaga[1,3,4], Nitesh V. Chawla[1,2,3,4]

[1]Department of Computer Science and Engineering, University of Notre Dame, IN 46556, USA
[2]Data, Inference, Analytics and Learning Lab(DIAL), University of Notre Dame, IN 46556, USA
[3]Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame, IN 46556, USA
[4]University of Notre Dame, IN 46556, USA

INTRODUCTION

Twitter is a micro-blogging site which enables individuals to write about their daily activities, express opinions, share information and connect with other users and businesses. Twitter is widely used by companies or brands to reach out to their customers, increase awareness about various topics and products.

According to a Twitter report[1], there are 284 million monthly active users, and 500 million tweets shared every day. This provides a unique business opportunity for companies to reach out to their customers, increase awareness and interaction among users about a topic or product and keep them engaged with their content. Using the 80/20 principle[2], 80% of the Tweets should be directed towards interacting and engaging with the current followers.

The relevant information cannot be captured using only the hashtags, mentions or keyword search, we need to model user's interests based on their tweeting patterns and the content of the Tweet. The companies need to listen to the trending topics amongst their followers to connect with them. In order to discover content a user is or might be interested in, we need to build a topic profile for each user which contains high level topics which he/she discusses in the post[3].

PROBLEM STATEMENT

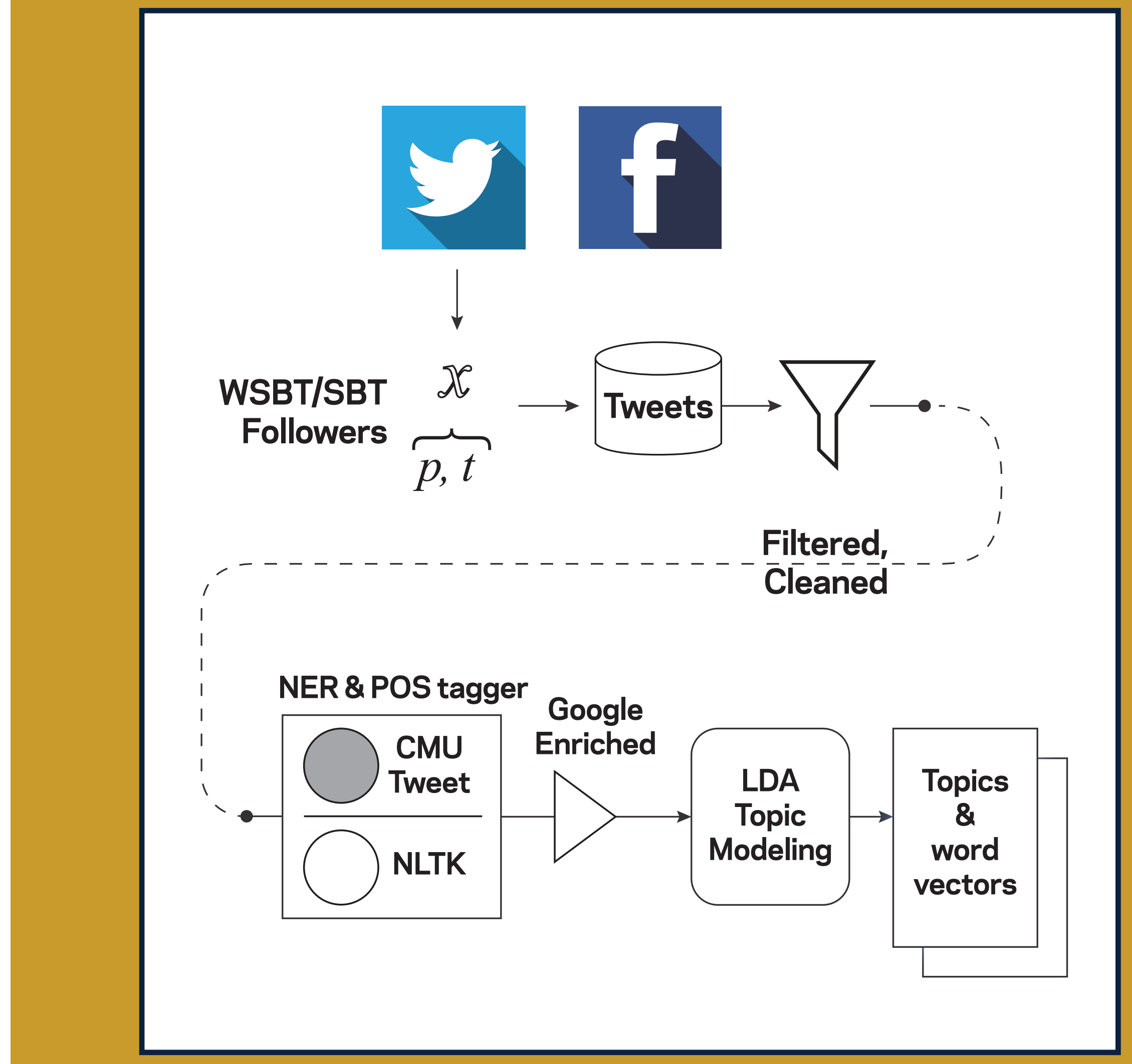
To identify the topics the followers are most interested in and to bridge the gap between the content published by media companies and what followers want to read.

DATA

We worked with the Twitter followers of a local media company - Schurz Communications Inc.[4] which own the South Bend Tribune (SBT) and WSBT-TV. The dataset consisted of tweets by their followers divided into two categories: 3700 tweets from prolific tweeters and 2500 tweets from typical tweeters.

The company defined prolific tweeters as individuals that tweet/retweet content and have a network of followers significantly more when compared to a typical follower.

FRAMEWORK



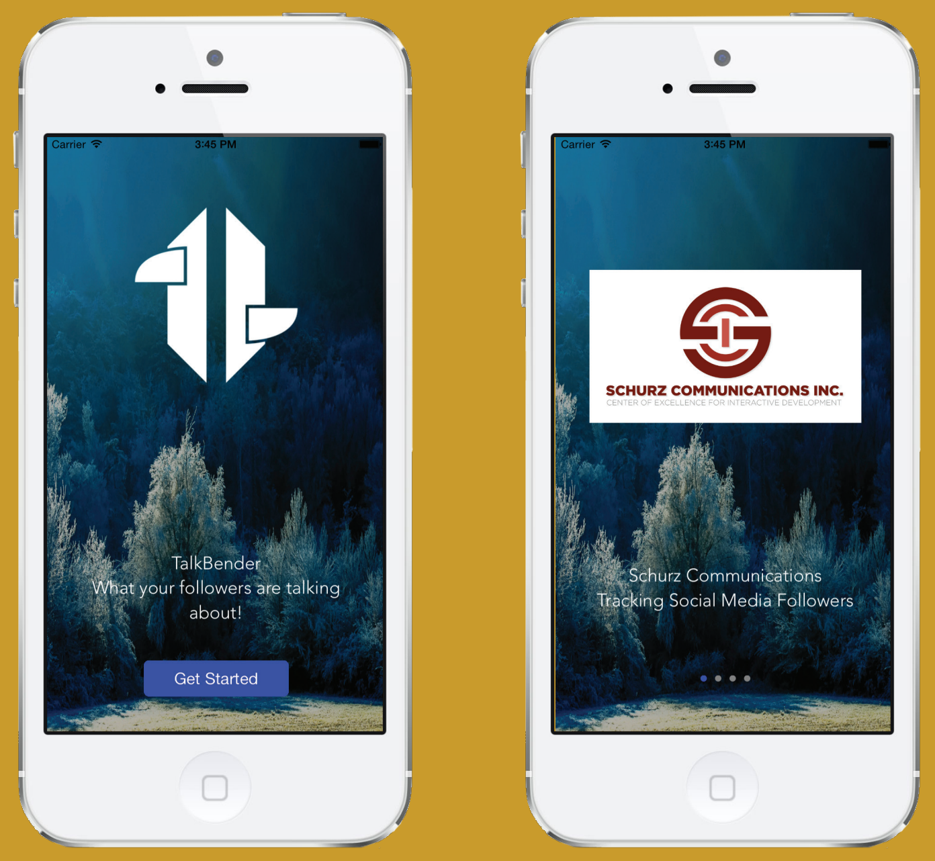
Given the limited text (140 characters) provided in each tweet, it becomes challenging to identify the topic in a tweet using the topic models. To overcome, this problem we present a framework that enriches the tweets with open source data to apply existing topic models.

We leverage the search engine's ability to augment our limited knowledge with the database of millions of documents. After performing pre-processing on our data set, we extracted the nouns of the tweet to build a query. The query is then fed into Google Search Engine[5] to retrieve the top 20 documents.

After retrieving at most 20 documents for each tweet, we crawl each URL to get the content of that page. After performing basic preprocessing on this content, we perform feature selection using TF-IDF score. Therefore, we are no longer restricted by the limited text in the tweet. We then performed Latent Dirichlet Allocation[6] (LDA) on the document corpus to get topics.

VISUALIZATION

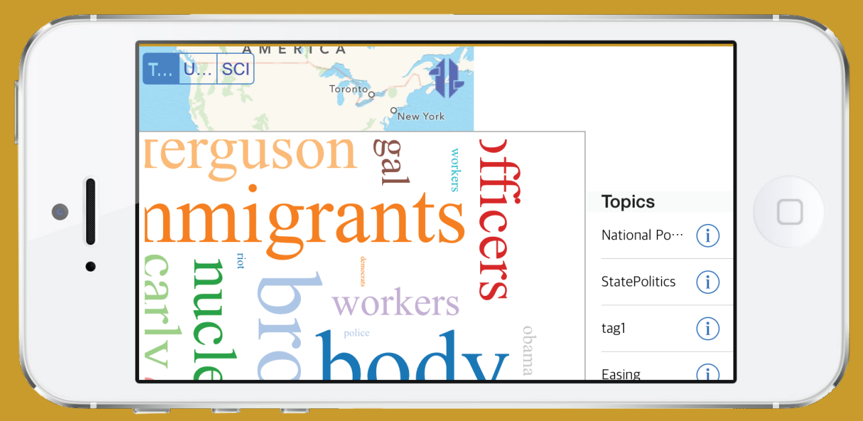
Since Twitter is a real-time information network, we built a mobile application which enables the companies to see the trending topics in real time. We also display the different topics a tweet is associated with and build a topic profile for the users. The application also takes in location and time to observe more detailed patterns.


Carrier
3:45 PM
SCHURZ COMMUNICATIONS INC.
CENTER OF EXCELLENCE FOR INTERACTIVE DEVELOPMENT
Get Started

TalkBender
What your followers are talking about!

Schurz Communications
Tracking Social Media Followers

Mobile application customized for Schurz Communications Inc.



Visualizing topic profile of a user using a word cloud for easy navigation.

PRE-PROCESSING

Tweets were cleaned of its unnecessary characters such as emoticons using regular expressions to extract the desired text. We began by discovering the named entities in a tweet using the process of Named Entity Recognition (NER)[7]. We focused on the Parts-Of-Speech (POS) tagger[8] to extract the nouns from the tweet. We used CMU tweet tagger[9] which has been trained over a set of tweets.

Tweet:
'I have completed the quest 'Another Try' in the #Android game The Tribez. http://t.co/uHpfiTNi57 #androidgames, #gameinsight'

Preprocessed and 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" , ' J J') , (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')]

ANALYSIS

To identify the trending topic amongst the followers, we applied the framework on the tweets of the typical tweeters. We extracted the top 10 topics and list the words that constitute the topic with respective probabilities. Using such distributions, we built a topic profile for each user. We applied the same framework to the tweets coming from the company's Twitter page and identified the difference in topics. The correctness of the topics assigned to the tweets have been manually verified.

Captures two topics obtained over the dataset by running LDA: World Politics and Ferguson Unrest. [5]

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		

[1] Twitter Internal data,2014 | [2] https://business.twitter.com/basics/ how-to-create-a-twitter-content-strategy | [3] Zhao W. X., et. al. "Topical keyphrase extraction from Twitter" ACL 2011 | [4] www.schurz. com | [5] www.google.com | [6] Blei D. M., et. al. "Latent Dirichlet Allocation" Journal of Machine Learning 2003 | [7] Ritter A., et. al. "Named Entity Recognition in tweets: an experimental study" EMNLP 2011 | [8] Toutanova K., "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger" EMNLP 2000 | [9] Gimpel K., et. al. "Part-of-speech tagging for Twitter: annotation, features, and experiments" HLT 2011.