Detecting Communities by Sentiment Analysis of Controversial Topics

SBP-BRiMS 2016

Kangwon Seo[1], Rong Pan[1], & Aleksey Panasyuk[2]

[1]Arizona State University

[2]Air Force Research Lab

July 1, 2016

Outline

I. Introduction

II. Overview of Dataset

III. Methodology

IV. Phase I : Topic Extraction

V. Phase II : Sentiment based Community Detection

VI. Conclusion

INTRODUCTION

- Using social media to detect and monitor emerging phenomena has attracted more and more attention from social scientists, political scientists, military strategists, etc.

- Twitter is an ideal source for collecting information of public opinions and public sentiments. With less than 140 characters, each tweet is more likely to express a Twitter user's feeling towards a particular event.

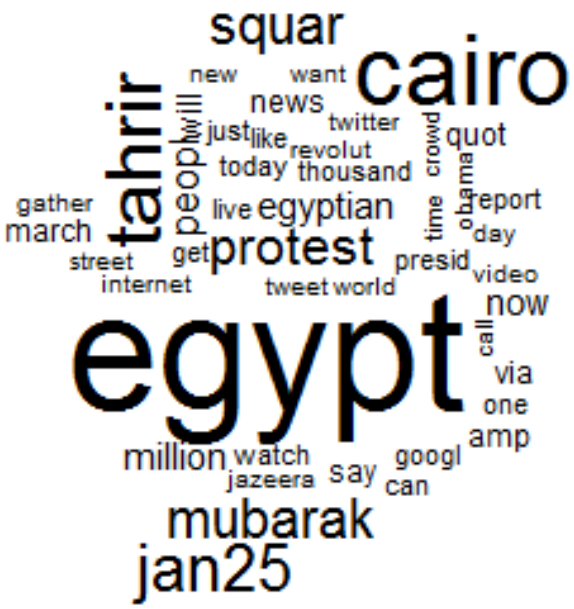
- There has been less attention of research to use sentiment analysis for online community detection and monitoring.

- In this research, we demonstrate a practical approach to social network analysis for 1) extracting a controversial topic from Twitter messages; and 2) identifying distinct online communities and prominent nodes in these communities.

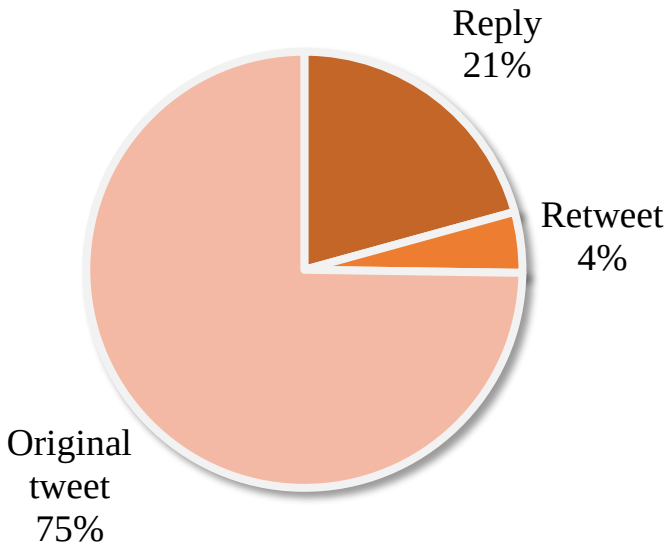
- In specific, we analyze Twitter data during the period of the Egyptian Revolution of 2011.

Overview of Dataset

- A collection of tweets (950K tweets with id, language, text message, user name, etc.) about Egypt during 19 days (Feb. 1, 2011 ~ Feb. 19, 2011) provided by Army Research Lab (ARL)
- We extracted tweets written in English (620K tweets).



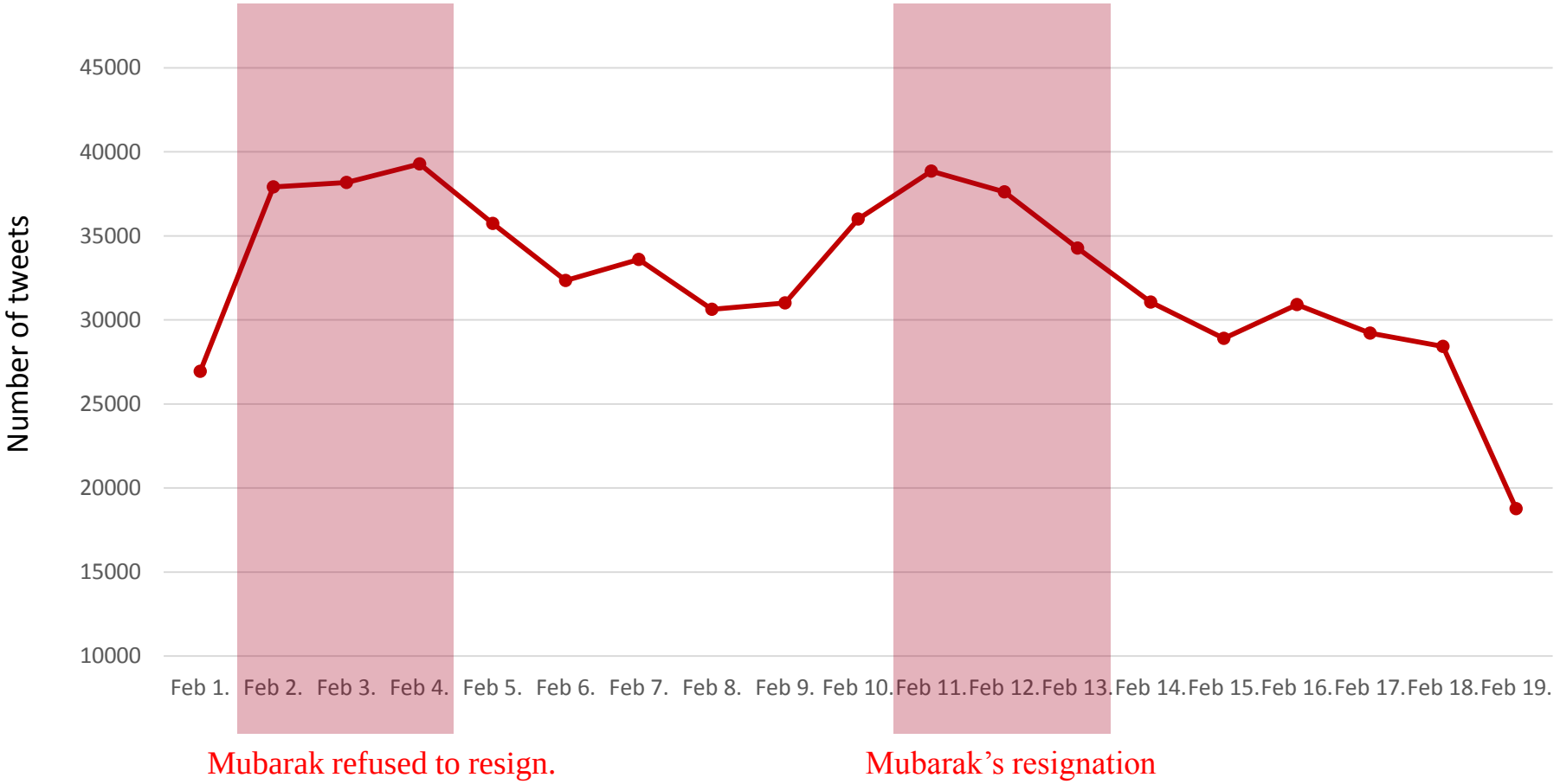
Word cloud with most frequent terms



Message types

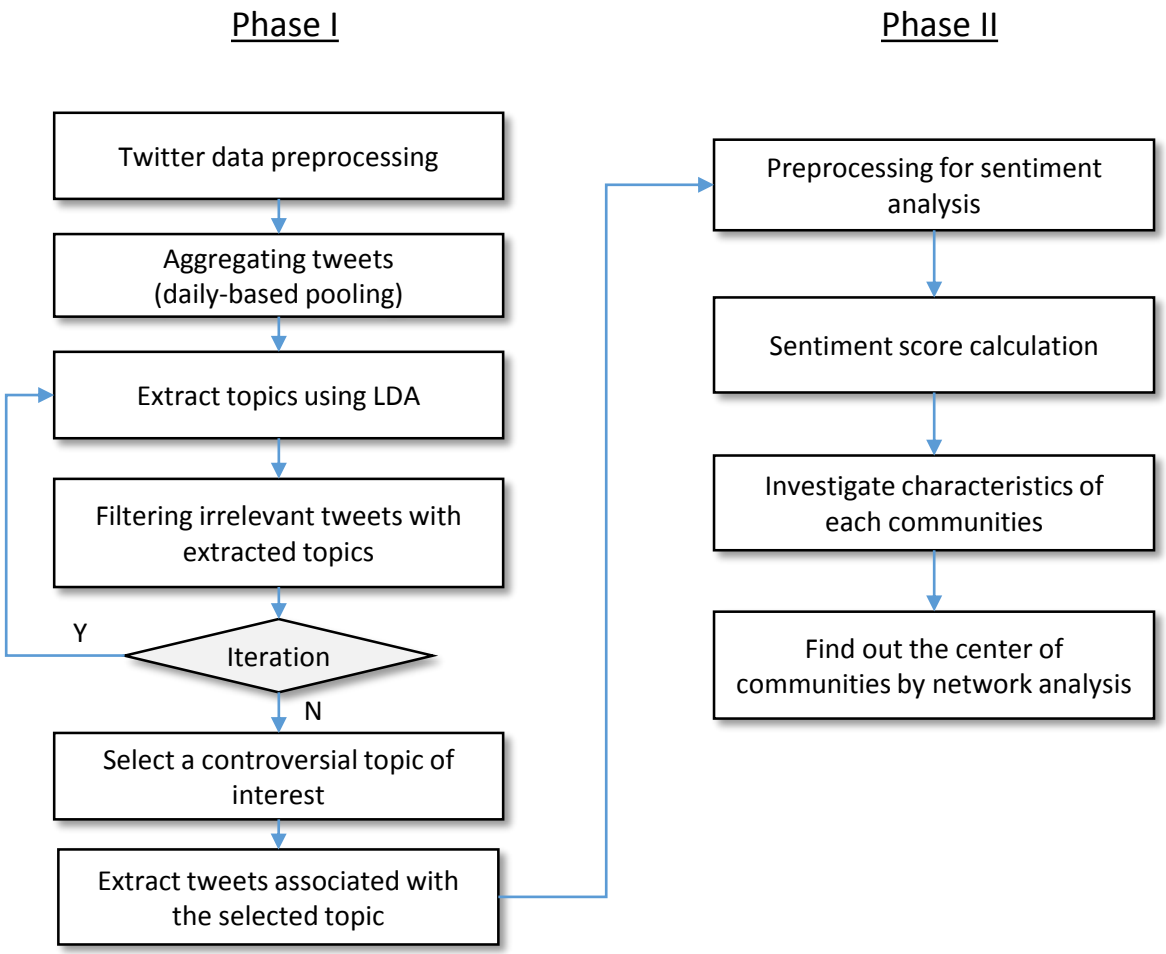
Overview of Dataset

- The number of tweets for each day peaked around Feb. 2 and Feb. 11, which corresponds to Mubarak's refusal to resign and Mubarak's resignation, respectively.



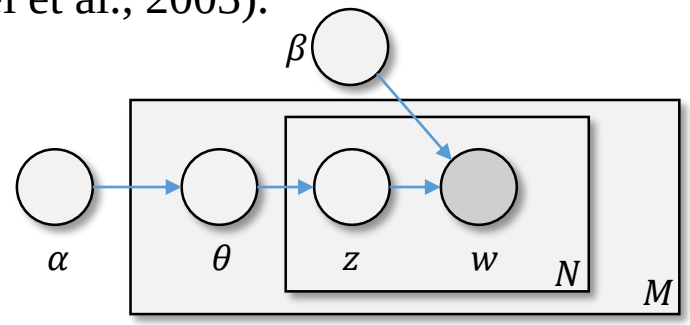
METHODOLOGY

- In phase I, a controversial topic of interest is identified and irrelevant tweets to this topic are filtered out.
- In phase II, sentiment analysis is performed and communities are identified.

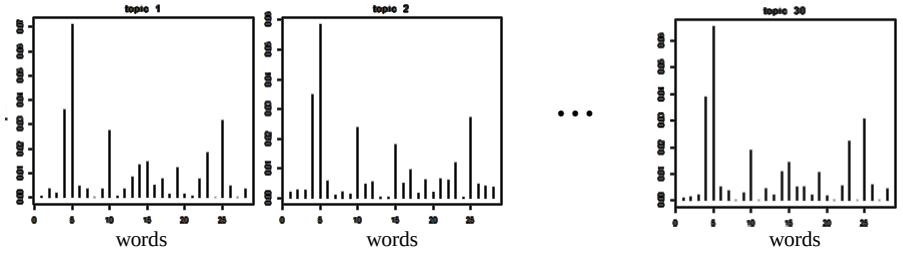


PHASE-I : TOPIC EXTRACTION

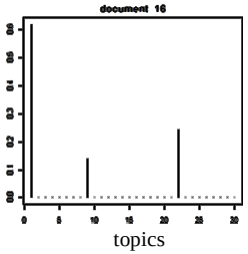
- Latent Dirichlet Allocation (LDA) is a flexible probabilistic generative model for topic modeling (Blei et al., 2003).



- According to LDA, a document is generated by following steps:
  - The term (word) distribution β for each topic is determined.



  - The topic proportion θ for a document is chosen from $Dirichlet(\alpha)$



  - Choose a topic Z from $Multinomial(\theta)$ for each word in a document.
  - Choose a word W from a multinomial probability distribution given the topic Z.

PHASE-I : TOPIC EXTRACTION

- One challenge of applying LDA on twitter messages is that a single twitter message is too short (< 140 characters) to be used as a document.

Use aggregated Twitter messages	Modify the standard LDA model
➢ Hong & Davison (2010) ➢ Mehrotra et al (2013) ➢ Author-based, Hashtag-based pooling ➢ Temporal pooling	➢ Zhao et al (2012) : Twitter-LDA model ➢ Ramage et al (2010) : labled-LDA

- In this research, we used daily-based pooling scheme (tweets posted on the same day are aggregated to a document) without modification of the standard LDA.

PHASE-I : TOPIC EXTRACTION

- We chose to fit 30 topics by LDA model.

Table 1. Extracted topics(part) by LDA

1	5	8	12	15	19	23	28	30
egypt	egypt	egypt	egypt	egypt	egypt	cairo	egypt	egypt
cairo	cairo	cairo	cairo	cbs	libya	egypt	cairo	cairo
bahrain	tahrir	jan25	tahrir	logan	jan25	jan25	tahrir	mubarak
revolution	can	google	will	lara	tahrir	revolution	jan25	tahrir
tahrir	video	protests	jan25	cairo	news	like	square	people
people	will	tahrir	like	news	square	news	news	square

- Most topics are dominated by a few similar terms (“Egypt”, “tahrir”, “jan25”, “cairo”); while some terms are relatively unique (“cbs”, “logan”, “lara” in topic 15).

- We also found that the document of each day involves only one or two topics, and hence we could determine the topical subjects of each day by looking at the highly ranked words in corresponding topics.

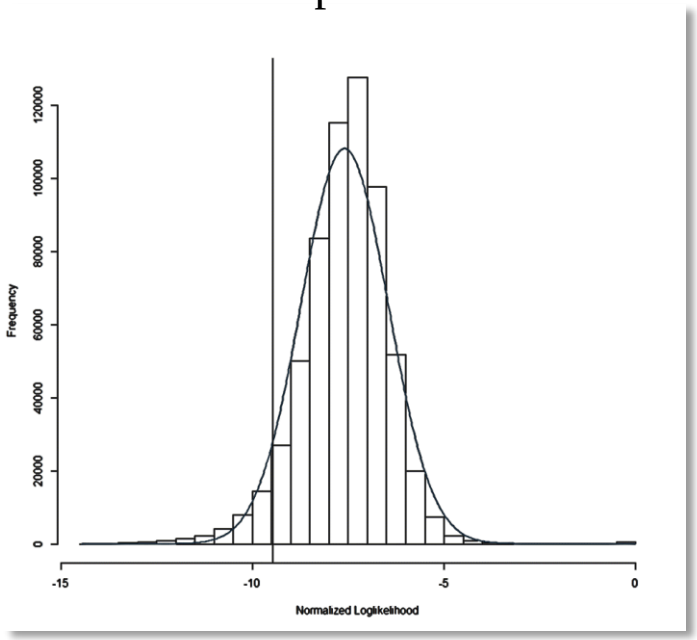
PHASE-I : TOPIC EXTRACTION

- To filter out Twitter messages which may not be related to any topic extracted from LDA, we fit the individual tweets (not aggregated tweets at this time) to LDA model once again; and obtained the topic proportions θ for each tweet.

- We calculate the following measure to determine relevance to the topics from LDA of each tweet.

$$relevance(T) = \frac{\sum_{i=1}^{n} \log p(w_i)}{n}$$

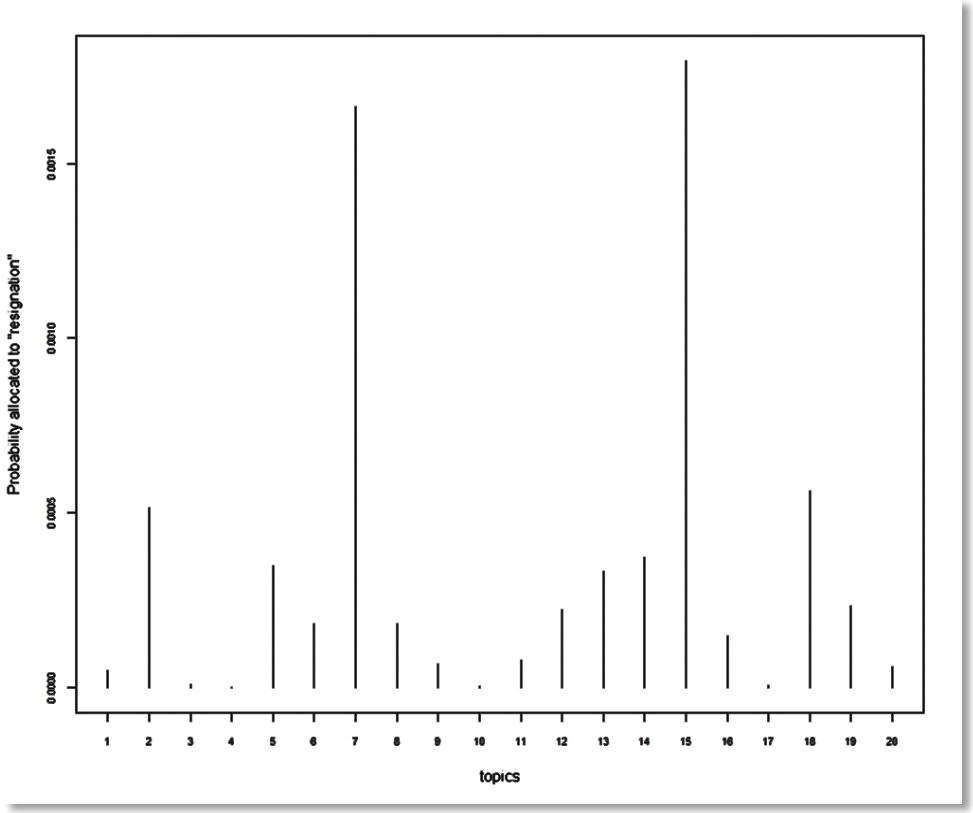
- We removed tweets which have *relevance* smaller than the 5% quantile of the normal distribution fitted to *relevance* values.



- We repeated the topic modeling on the remaining Twitter data with a smaller number of topics (20 topics); then removed irrelevant tweets using *relevance*.

PHASE-I : TOPIC EXTRACTION

- After 2nd iteration of topic modeling, we chose "the resignation of Mubarak" as the controversial topic of interest.



Probability of the term "resignation" for each topic

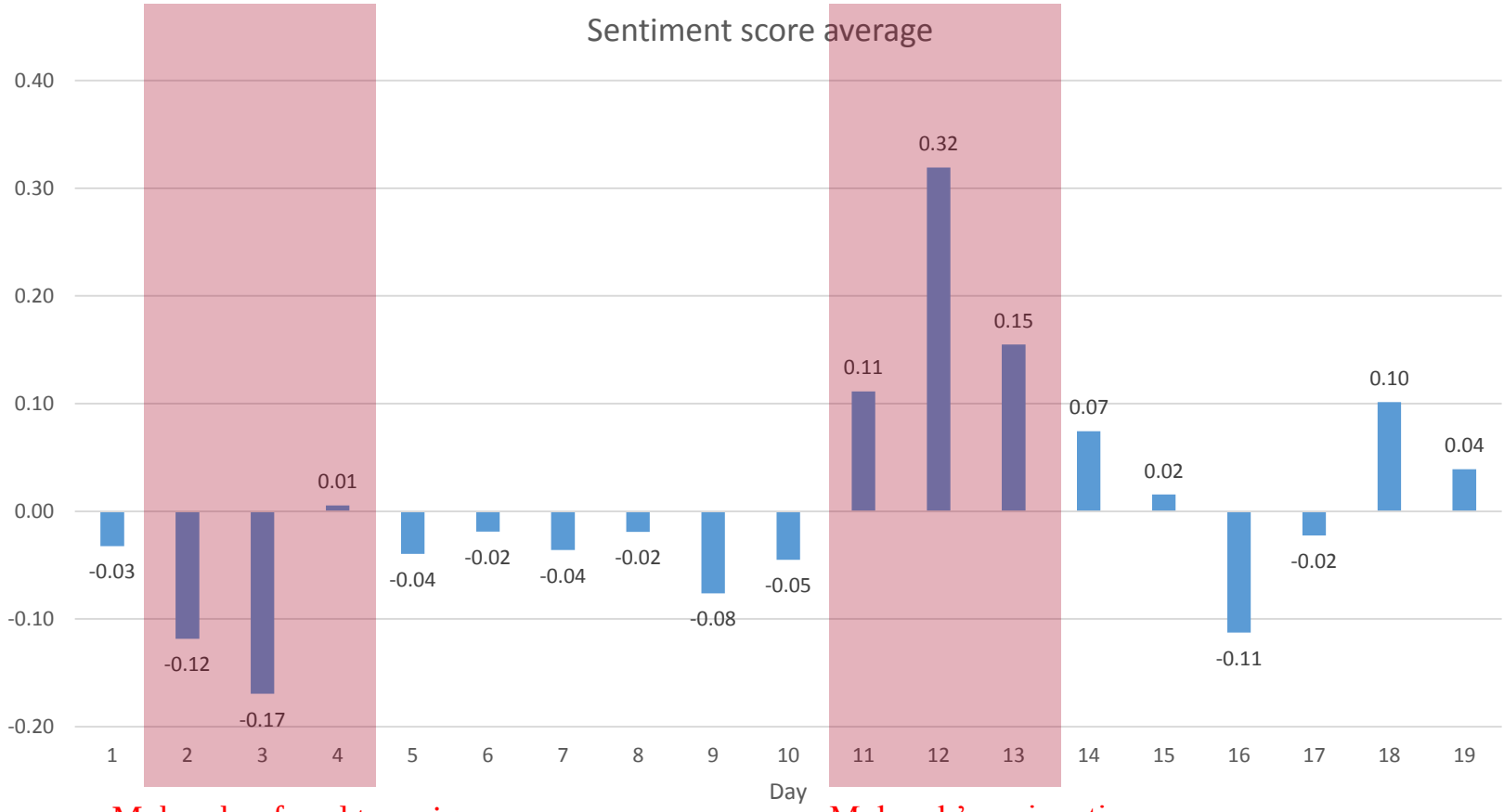
- Topic 7 and Topic 15 are thought to be highly related to the resignation of Mubarak since those two topics allocate relatively high probability on the term "resignation", comparing to other topics.
- We extracted tweets that had the largest weights on Topic 7 or Topic 15. (27K tweets)

Phase-II : Sentiment-based Community Detection

- With the selected topics from twitter corpus, we assess the twitter user's attitude (positive, negative, or neutral) to those topics.
- To quantify the polarity in a given text, we compared the text with known lexicons of positive and negative words. We used Bing Liu's sentiment lexicon for this task.
- Based on the sentiment score, we partitioned users to different communities.
- We also exploit tweet-retweet and tweet-reply networks to identify the users who play a role of opinion leader in the community.

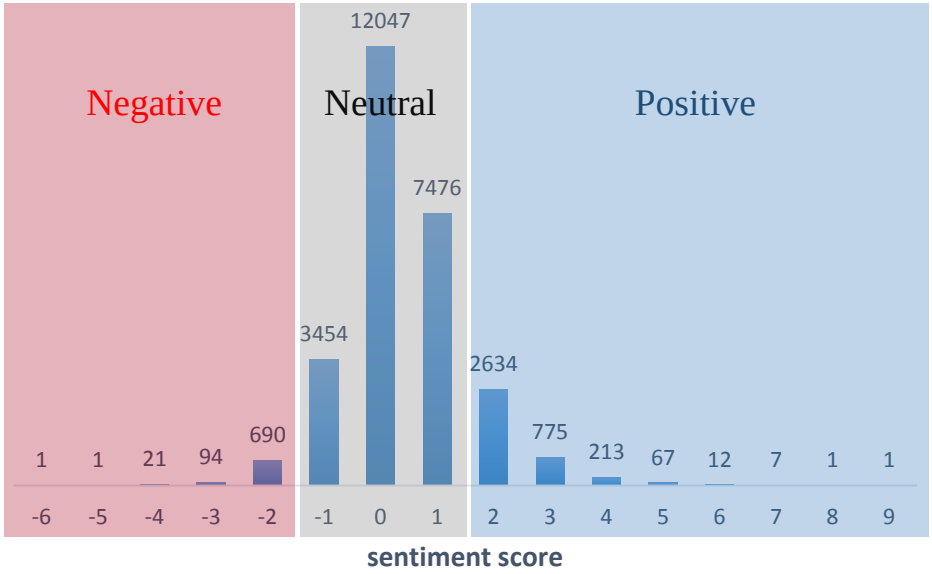
PHASE-II : SENTIMENT-BASED COMMUNITY DETECTION

- The sentiment scores of each tweet were calculated by a simple voting between the number of positive words and negative words in the tweet.



PHASE-II : SENTIMENT-BASED COMMUNITY DETECTION

- Treating the sentiment scores -1, 0 and 1 as neutral, we used the tweets with scores higher than 1 or lower than -1 to construct two opposite communities.
- We also investigated the magnitudes of 6 emotional items in each community.



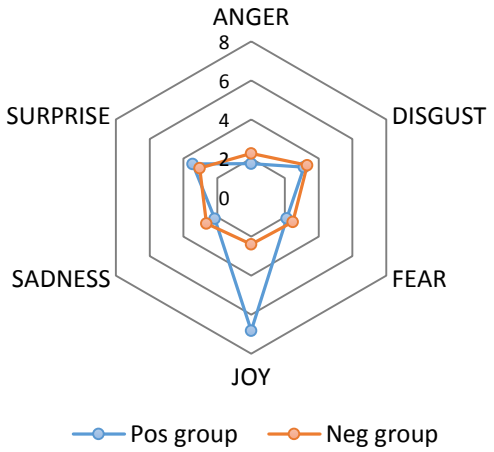
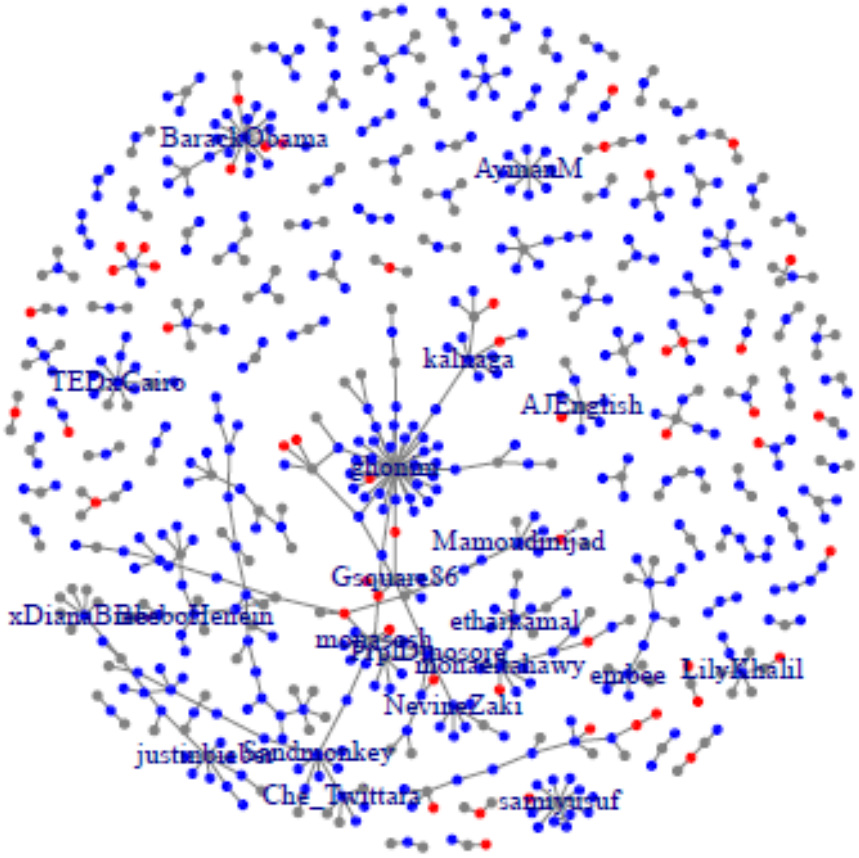


Table 2. Summary of sentiment & emotion analysis

	Sentiment			Emotion					
	users	tweets	score	anger	disgust	fear	joy	sadness	surprise
Pos.	2,978	3,831	2.36	1.74	3.12	2.11	6.82	2.15	3.46
Neg.	525	635	-1.98	2.27	3.32	2.47	2.40	2.64	3.04

PHASE-II : SENTIMENT-BASED COMMUNITY DETECTION

- The network structure of tweets provides useful information of how Twitter users communicate each other.
- We constructed a network from all pairs of users who were linked by tweet-retweet and tweet-reply.



- "ghonim" is likely to be the opinion leader in this network, which is the username of Wael Ghonim who had led anti-government protesters during Egypt revolution.
- "BarackObama" is the username of the U.S. president. He expressed welcome at Mubarak's decision to step down.
- "samiyusuf" is the username of Sami Yusuf, who is a British Muslim singer-songwriter and released a song prompted by Egypt revolution.

Conclusion

- By detecting and analyzing communities formed around a controversial topic we can paint a picture of how polarized a country is. This knowledge is valuable for predicting some future events such as violent civil conflicts.

- In this research, we made use of Twitter data to identify a controversial topic of interest and the sentiment / emotions expressed towards the topic.

- In addition, the network structure of these communities was utilized to reveal Twitter users who played important roles in their respective communities.

Key References

1. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993-1022 (2003)
2. Hong, L., Davison, B. D.: Empirical Study of Topic Modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics, pp. 80-88. ACM (2010)
3. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 889-892. ACM (2013)
4. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., Li, X.: Comparing Twitter and Traditional Media Using Topic Models. In: Advances in Information Retrieval. pp. 338-349. Springer, Berlin Heidelberg. (2011)
5. Ramage, D., Dumais, S. T., Liebling, D. J.: Characterizing Microblogs with Topic Models. ICWSM, 10, 1-1. (2010)
6. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM (2004)
7. Feinerer, I., Meyer, D. Hornik, K.: Text Mining Infrastructure in R. J. Stat. Softw. 25(5), 1-54 (2008)
8. Grun, B., Hornik, K.: topicmodels: An R package for Fitting Topic Models. J. Stat. Softw. 40(13), 1-30 (2011)
9. Marwick, B.: Discovery of Emergent Issues and Controversies in Anthropology Using Text Mining, Topic Modeling, and Social Network Analysis of Microblog Content. In: Data Mining Applications with R, p. 514. Academic Press, New York (2014)

Thank you.
Q & A