

Does Rhetoric Suggest a Social Conflict?

Masoud Makrehchi
Faculty of Engineering and Applied Science
University of Ontario Institute of Technology
Oshawa, ON, Canada
masoud.makrehchi@uoit.ca

ABSTRACT

In a polarized society, rhetorical arguments are usually expressed by strong, extreme terms which by themselves carry a positive or negative sentiment about one side of the social debate or conflict. By detecting extreme terms in a social-political text such as a blog post, we are able to automatically detect the sentiment of the text about polarizing issues in a divided society. On the other hand, during social and political conflicts in polarized societies, we can observe a shift from mainstream language to extreme language and rhetoric. In this research, we introduce language gap which is the distance between mainstream and rhetoric. Then we illustrate that there is a correlation between language gap and social conflicts.

1. INTRODUCTION

By monitoring societies that are traditionally polarized over certain issues, for example in the United States over abortion or gun right, one can easily observe not only the society but also the vocabulary is polarized. The polarized discourse, although, is the result of deep disagreement about controversial issues, it also helps to heat up the debate and widen the gap between two sides of the discussion[4]. In other words, a polarized vocabulary, along with other facts, can be used as an indicator to detect if a debate is turning into a crisis or it is fading and becoming more the language of extremists in both sides. In the later case, mainstream language is being used by the majority of debaters while in the former case, instead of a constructive discussion, both sides are involved in accusation, tagging, and labeling the other side. In this paper, a qualitative and analytic approach is proposed to measure the depth of a social-political crisis using the polarized vocabulary which is called Language Gap.

One of the most recent events that demonstrated deep dividedness in a society was of the 2009 Iran presidential election. The outcome of the election was disputed by the opposition and caused a prolonged, deep political crisis. The election itself, presidential debates, post-election demonstrations, and the chain of events long af-

ter election have been widely reflected in Iran's Farsi Blogosphere which is one of the largest in its kind. In order to support the hypothesis, Farsi blogs were monitored during 2009 and about 190 Farsi topics trends used by both opposition and pro-government blogs are were collected[2, 3].

Iran, with more than 700,000 Farsi blogs, is one of the top ten countries in the number of bloggers in the world[1]. They are mostly based in Iran, although there are many blogs written by Iranians in English and other languages inside and outside the country. The Farsi language is ranked as the second-most-popular language in the entire blogosphere.

Topic time series are very noisy. They are smoothed using a simple moving average smoothing function by which every smoothed observation is simply the mean of the last w observations. The parameter w is called smoothing window. Let x_t be the original observation at time t and y_t be its smoothed version:

$$y_t = \frac{1}{w} \sum_{k=0}^{w-1} x_{t-k} \quad (1)$$

Representing topics as time series, Latent Semantic Indexing (LSI) is employed to extract semantic relations between topics and time intervals which can be grouped as events. What we achieve, by applying LSI, is effective dimensionality reduction and in the same time, extracting more relevant features.

2. TOPIC VS. EVENT

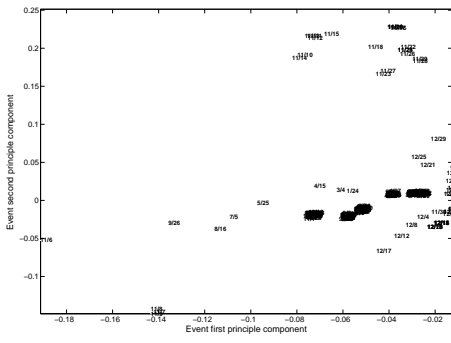
In both topics and events (days), some points are noise and should be removed. In Fig. 1 topics and event are visualized using first two principle components before filtering. Since topics are collected from blogs using the criterion of high frequency, it is possible that many noise topics are added to the set which are not relevant to the conflict under study.

Using density-based clustering both events and topics are filtered to identify densely populated topics and events and remove those which are sparse. While density-based clustering is used to first filter events and then cluster them into event clusters, we only apply the clustering algorithm to filter topics because topics are manually categorized into two extreme and mainstream topics.

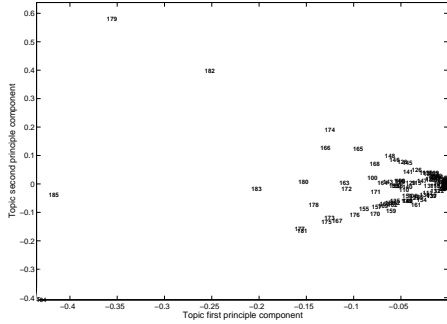
After filtering both days and topics and clustering days to find a structure among them (Event Clusters), both day and topic points are mapped in a same space. Fig. 2 illustrates the mapping into two dimensional space for better visualization. According to the figure, blue triangles (mainstream topics) are relatively distant from event clusters which represent dense events and conflicts. On the other hand, extreme topics are close to the conflicts compared to the mainstream topics. Although we cannot draw a conclusion on causal relations between rhetorical language and social-political

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.



(a)



(b)

Figure 1: Scatter plot of two first principle components of (a) events and (b) topics.

Table 1: Accuracy of detecting extreme topics from mainstream topics.

Classifier	SVD	Event-based representation using DBSCAN		
		$MinPnts = 2$	$MinPnts = 4$	$MinPnts = 6$
Linear	0.6618	0.7786	0.7302	0.7469
SVM	0.7589	0.7908	0.7849	0.7718

conflicts, we can observe extreme language as a symptom and indicator for an emerging conflict. The conclusion is interesting since by mining the language used in social media, we can build models to predict the likelihood of a social-political conflict.

Using a classifier, the predictability of conflicts by monitoring rhetorical language in polarized society is demonstrated. Each topic is represented by its distance from events or conflicts (see Fig. 2). In other words, distances are considered as new features describing the topics. While classifying topic time series to distinguish rhetoric from main stream topics offers low accuracy, new distance-based features is able to classify two groups of topics with 80% accuracy (see Tab. 1).

This observation can be also explained by the social polarization concept. If we assume a deep conflict arises when the society is deeply divided and polarized, then we can expect a polarized vocabulary and a gap in language in social debates.

In order to estimate the correlation between the distance between the two groups of topics (rhetoric and mainstream) and events, the above mentioned experiment is conducted on the data set in monthly basis. In each snapshot, the normalized distance of each topic group from conflict events are estimated. Let t_i be a topic and

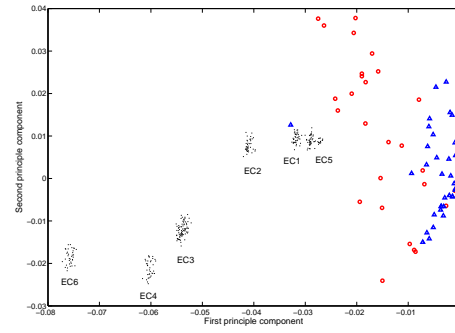


Figure 2: Event and topic clusters after removing outliers and noise by density-based clustering. Red circles are extreme (right and left) and blue triangles are mainstream topics.

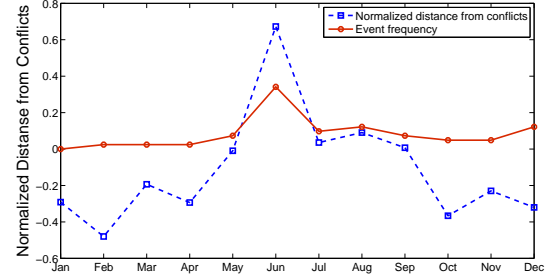


Figure 3: The correlation between normalized distance and the magnitude of social conflicts.

e_j be event j

$$f = \sum_{i=1}^n \sum_{j=1}^m dist(t_i, e_j) \quad (2)$$

Normalized distance d is estimated as follows:

$$d = \frac{f^+ - f^-}{f^+ + f^- + 1} \quad (3)$$

where f^+ (f^-) is the distance between mainstream (extreme) topics and the events. Fig. 3 demonstrates the predictability of the conflicts (represented by event frequency which is the number of conflict events per month) using normalized distance of topics and events.

3. REFERENCES

- [1] N. Alavi. *We are Iran*. Soft Skull Press, 2005.
- [2] D. Gaffney. *iranElection : Quantifying Online Activism*, pages 1–8. 2010.
- [3] L. Khalili Gheidary. Social media and Iran’s post-election crisis. In *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management*, Seattle, WA, USA, September 2009.
- [4] B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, Sept. 2008.