Analyzing the Political Sentiment of Tweets in Farsi

Elham Vaziripour and **Christophe Giraud-Carrier** and **Daniel Zappala**
Brigham Young University
Computer Science Department
Provo, UT 84602
Email: elham@internet.byu.edu, cgc@cs.byu.edu zappala@cs.byu.edu

Abstract

We examine the question of whether we can automatically classify the sentiment of individual tweets in Farsi, to determine their changing sentiments over time toward a number of trending political topics. Examining tweets in Farsi adds challenges such as the lack of a sentiment lexicon and part-of-speech taggers, frequent use of colloquial words, and unique orthography and morphology characteristics. We have collected over 1 million Tweets on political topics in the Farsi language, with an annotated data set of over 3,000 tweets. We find that an SVM classifier with Brown clustering for feature selection yields a median accuracy of 56% and accuracy as high as 70%. We use this classifier to track dynamic sentiment during a key period of Irans negotiations over its nuclear program.

Introduction

Our focus in this paper is on the extent to which Twitter data can be mined to discover the opinions of Iranians toward various political topics. Many citizens may be reluctant to share their true opinions when contacted by a stranger for a telephone poll, especially on politically sensitive topics. Twitter has the potential to act as an alternative measure of political opinions due to its widespread use and public nature. Indeed, there is some evidence demonstrating that sentiment word frequencies in Twitter messages have correlations as high as 80% with surveys on consumer confidence and political opinions (O'Connor et al. 2010). One challenge with political topics is that Twitter users may skew toward those who are disaffected, living abroad, or toward certain demographics. Our work represents a first attempt at understanding political sentiment among Farsi-speakers so that we can begin to address these challenges.

The Farsi language, also known as Persian, is from the Indo-European family of languages. It is written from right to left and has been greatly influenced by the Arabic language, so much so that it has the same alphabet but with four additional letters. Farsi introduces additional challenges for sentiment analysis, due to a lack of linguistic tools. First, most of the automatic text classification systems used with success on English texts (e.g. (Pennebaker et al. 2007)) are not useful for Persian. This is mainly due to a lack of a comprehensive Persian WordNet, subjectivity lexicon for informal text and challenges concerning the nature of the Persian language (Saraee and Bagheri 2013). Although some attempts have been made to make a Persian version of WordNet for formal Farsi words, known as FarsiNet, no publicly available product has been introduced yet Second, using colloquial words is very popular with Persian-speaking Twitter users, which can obscure meaning. Third, orthography in Persian includes hidden diacritics, hidden possessive morpheme, the multiple shape of affixes, no gender distinctions, using pseudo space in the internal structure of some words and no punctuation in writing. Finally, where morphology is concerned, the Persian language is full of compound verbs, the parts of which can have long-distance dependency. That is to say, other words can be present between the non-verbal element and light verb. Farsi is a free word order language with complicated syntactic trees. Therefore, not only do adverbs appear everywhere in the sentence, but adjectives also follow or precede the nouns .

In this study, we investigate political discussions about Iran, the United States, and the U.N. sanctions during the time period of February 2013 to December 2013. We have collected over 1 million tweets using Twitter's streaming API, using a set of Farsi keywords to target selected topics. Using an SVM classifier with features identified by Brown clustering, we are able to achieve a median accuracy of 56%, compared to a baseline of 46%, with one week having accuracy as high as 70%, given a baseline of 58%. We use this classifier to track dynamic sentiment as expressed by Twitter users during a key period in Iran's negotiations with its nuclear program, and show that changes in sentiment correlate with key words found with LDA topic modelling.

Related Work

Two works are particularly relevant to our topic of classifying the political sentiment of individual tweets. Wang et al. have created a system to identify real-time voter sentiment, classifying tweets as positive, neutral, negative, or unsure with an accuracy of 59% (Wang et al. 2012). This work uses unigram features with a naive Bayes classifier and Mechanical Turk for providing labels of training data. Bakliwal studies the problem of finding an accurate classifier for political tweets and is able to achieve 61% accuracy using an SVM

classifier with unigram features together with some emoticons, URLs and hashtags (Bakliwal et al. 2013).

Other works on topical sentiment analysis are able to achieve higher accuracy due to characteristics of the data set being analyzed. For example, an early work on sentiment classification of movie reviews from a Usenet newsgroup achieved an accuracy of nearly 83% with an SVM classifier and unigram features (Pang, Lee, and Vaithyanathan 2002). Users are likely to be more forthright in their rating of movies as compared to discussions of sensitive political topics in the Middle East.

Some research has shown that using Twitter-specific features helps classify the sentiment of individual tweets. Bakliwal et al. have been able to achieve an accuracy of 88% on a corpus of Tweets that have been labeled based on emoticons (Bakliwal et al. 2012). Their preprocessing steps include removing stop words, stemming and noun identification with WordNet. Other work has classified sentiment toward trending hashtags (Wang et al. 2011). We explored these methods, but our experience with our data set suggests that hashtags do not indicate sentiment toward a topic, and emoticons are rarely used.

A large body of research examines political orientation of users, as opposed to sentiment toward individual political topics. Here, the research is split on whether to use the content of tweets or the relationships among users. Golbeck and Hansen compute political preference based on the average of the political scores of political figures that a user follows, with the scores taken from an external ranking source (Golbeck and Hansen 2011). In contrast, Zamal et al. use a large number of features, such as k-top words, k-top stems, frequency statistics, and retweeting tendency, along with an SVM classifier, to compute political orientation as Republican or Democrat (Al Zamal, Liu, and Ruths 2012). They also show that including neighbor attributes improves classification accuracy. Unfortunately, many of these methods have been applied to classifying politicians or news sources, as opposed to more ordinary users. Cohen and Ruths have demonstrated that these methods, which typically report accuracy as high as 90%, achieve only 65% accuracy for normal users (Cohen and Ruths 2013).

Methodology

We collect data using the Twitter Streaming API using the "statuses/filter" endpoint. This endpoint allows us to query the stream for keywords related to political topics. We use keywords in Farsi, then filter out only those tweets actually in the Farsi language, using the language field of the returned documents. (Some words are written the same in Farsi and Arabic.) Since our keywords are in Farsi, there is a very high likelihood that the tweets are from Iranians, living either inside the country or abroad. While collecting Twitter data we were never rate-limited, so we were able to receive all of the Tweets on these topics in Farsi.

Our data set was collected between the end of February 2013 and the beginning of December 2013 using the keywords (in Farsi) of John Kerry, Israel, America, and Obama for the "U.S government" topic, Khamenei, Internet, freedom, political, president, government, and Iran for the "Iranian government" topic and dollar, inflation, economic, and sanction for the "Sanctions" topic. These keywords were chosen based on manually viewing recent tweets on these topics to find commonly used words. We collected a total of 1,025,303 tweets, including 107,997 retweets, posted by 72,454 unique users during that time period.

One of the main challenges in preparing the training data for a classifier is to label it consistently. During an initial phase, we randomly selected 10 tweets from each topic – the Iranian government, the U.S. government, and the U.N. sanctions – and asked 9 Farsi speakers to label the sentiment of each tweet toward the topic. To label tweets toward the topic "Iranian government" and "U.S. government", we asked our helpers to label the attitude of the tweets toward the government. For the "Sanctions" topic we asked them to label a positive sentiment if the tweet expressed hope that the sanctions would be lifted or life would improve, while a negative sentiment indicated the people were suffering or felt unhappy with the sanctions. Each person labeled the tweets on a scale from 1 to 5. We then grouped the ratings into negative (1,2), neutral (3), and positive (4,5). We then selected five of the helpers whose labels were in strong agreement with the rest of the group to label more than 3000 tweets, selecting 100 random tweets for each week and topic. For the supervised classifiers we have used labels with the majority vote from annotators.

The labeled tweets are 35% neutral, 37% negative, and 27% positive. Although the distribution of labels is rather balanced, the distribution for each week is imbalanced.

Results

For our experimental results, we focus on the data we have collected for September 2013, because we have labeled training data for this month. This includes 274,032 tweets from 17,844 unique users. The corpus contains 4,573,601 words, with 355,541 unique words after stop words have been removed, and 319,160 unique stemmed words.

For most of our analysis, we divide September into four weeks and train a classifier on a particular topic for that week. We use the topics of Iran, the United States, and sanctions. In all cases where we train a classifier, we divide the labeled data into a training set (80%) and test set (20%). Each result is cross-validated with 10 trials.

Classifier Accuracy

Using Brown clustering for feature selection significantly improves the accuracy of a classifier for this task. Brown clustering is an unsupervised machine learning method for assigning words to classes based on the frequency of their co-occurrence with surrounding words in a large corpus (Brown et al. 1992) (Ushioda and Kawasaki 1996) (Miller, Guinness, and Zamanian 2004). This is an attractive method for feature selection, because we have more than a million tweets available in our corpus.

Brown clustering works by starting from a set of singleton classes and repeatedly merging the classes based on an average mutual information objective function until achieving a final number K of classes. By tracking the merging process

Week	Accuracy NB	Accuracy SVM	Fscore NB	Fscore SVM	Baseline
Iran					
1	0.62	0.65	0.71	0.71	0.53
2	0.52	0.56	0.69	0.67	0.46
3	0.44	0.43	0.52	0.47	0.4
4	0.43	0.47	0.48	0.61	0.36
USA					
1	0.57	0.51	0.67	0.56	0.39
2	0.51	0.53	0.65	0.65	0.45
3	0.6	0.55	0.53	0.55	0.47
4	0.52	0.57	0.46	0.66	0.48
Sanctions					
1	0.63	0.7	0.48	0.67	0.58
2	0.6	0.63	0.72	0.61	0.51
3	0.64	0.62	0.61	0.53	0.49
4	0.54	0.54	0.61	0.59	0.45
Median	0.55	0.56	0.59	0.61	0.46

Table 1: Accuracy of Naive Bayes and SVM classifiers with Brown clustering, 1000 clusters with cutoff 3

of clusters, we can create a hierarchical representation of the vocabulary. The resulting hierarchical clusters can be represented with bit-strings, which we can then use as features (Miller, Guinness, and Zamanian 2004).

The two parameters for clustering are the number of clusters and the cutoff word frequency, meaning the minimum number of times a word must appear in the corpus to be included in the clustering. We use our entire corpus from February to December to find clusters, with 100 and 1000 clusters of words and a cutoff of 3, 10, 20 and 40. We take our training and test sets from labeled data for each of the weeks in September 2013, on the topics of "Iran", "United States" and "Sanctions", resulting in 12 data sets (4 weeks and 3 topics). We found that 1000 clusters of words with a cutoff of 3 provides the best accuracy.

We tested both the Naive Bayes and SVM classifiers with Brown clustering, and found that it improved accuracy by 5% and 7%, respectively, as compared to using unigrams for features. Table 1 shows the accuracy and F-scores for each data set using Brown clustering with 1000 clusters and a cutoff of 3. The baseline performance is given by a classifier that simply chooses the majority classification each time (this is generally negative or neutral sentiment). The last row shows the median accuracy of the classifiers over all four weeks and all three topics. Brown Clustering improves on the baseline by 3% to 10%. For the best week, SVM has an accuracy of 70% as compared to a baseline of 58%.

To check for statistical significance, we bootstrap the confidence interval and check whether it overlaps. We find that for Naive Bayes, Brown clustering improves on unigrams with a 95% confidence interval. For SVM, Brown clustering improves on unigrams with an 85% confidence interval.

Dynamic Sentiment

We next use the SVM classifier with Brown clustering (1000 clusters, cutoff of 3) to examine how sentiment changes over time toward each of the topics. We train three SVM classifiers on the labeled data for the entire month of September 2013 on each of the topics separately. We then use the trained classifiers to score the sentiment of all tweets on each topic. Figure 1 shows how the sentiment changes for these topics during the sliding one-week period over the course of the month.

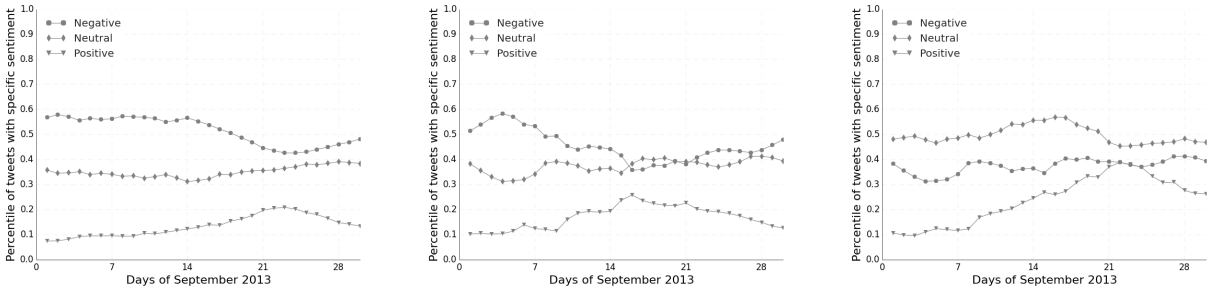
To correlate this sentiment to current events, we use LDA to discover the subtopics in the tweets during each week. We use the term *subtopic* to denote concepts found by LDA within each main topic. For example, in the "Sanctions" topic, we found subtopics such as "oil" and "medical". We configure LDA to find 10 subtopics, with 20 words in each subtopic. Table 2 shows one representative word (translated into English) for each subtopic identified by LDA during the third and fourth weeks, when sentiment was changing the most. Using these subtopics, we were able to identify several events that can help explain the dynamic sentiment during this month.

The biggest trend is clearly the greater positive sentiment toward the United States near the end of the month. At this time, the new Iranian President Rouhani attended the United Nations General Assembly in New York, with the intention of resolving the nuclear dispute through diplomatic negotiations. The ministers of the two countries met each other and Rouhani posted on his Twitter account about his willingness to arrive at an agreement in the Geneva meeting, which led to increased positive sentiment toward the United States during this time. We found key subtopics during the third week of September, such as "public" (referring to the assembly), "meeting", "negotiate", "flexible", "relation", "message", and "heroic". The term "heroic flexibility" is a term Seyed Ali Hosseini Khamenei used at the time to symbolize compromising.

There is likewise an increase in positive sentiment toward Iran at the same time, along with a decrease in negative sentiment. Prior to the UN assembly, Iran eased restrictions on Twitter and the Internet and released some political prisoners; we found this reflected in subtopics such as "filter", "lifting" and "freedom."

There is also an increase in positive sentiment toward sanctions and a decrease of negative sentiment during the middle of the month. We should be clear that this does not mean that Iranians felt the sanctions were good, but that they expressed hope that they might have some relief. During the third week we found subtopics such as "negotiation", "flexibility", "cheap", "sanction," and discussion of a lower cost of the dollar.

These results suggest that changes in sentiment regarding current events are manifested in political tweets in Farsi, indicating the potential for data mining of Twitter to act as a complementary method to traditional landline polling for finding opinions of large groups of people.



(a) Iran (b) Sanctions (c) United States

Figure 1: Dynamic sentiment during September 2013

Main topic	Week	10 LDA subtopics
Sanctions	3	revolutionary guards, drop, Roohani, gold coin, negotiation, flexibility, heroic, cheap, resonance, sanction
	4	down, drop, up, dollar, negotiate, trip, Hassan, auto makers, don't let, improvement
Iran	3	empty, nomads, lifting (of Internet filtering), filter, Internet, Islamic, shipping , freedom, prisoners, Geneva
	4	responsible, murder, boss, negotiations, Khatami, medal, wrestle, executions, protests, hejab
United States	3	Zarif, lock, dollar, meeting, public, Netanyahu, Moosavian, Khamenei, Middle east, Kerry
	4	speech, Roohani, Israel, flexibility, heroic, relation, message, English, dollar, reception

Table 2: LDA subtopics for ending weeks of September 2013

Conclusion

Although it can be argued that Iranian Twitter users are not necessarily an accurate representative sample of the whole population of Iran, our work represents a useful early step toward automatically classifying political sentiment among a large population. To provide a better alternative to traditional polling, additional work is needed to correlate the sentiment of individual tweets with the overall political opinions of a given person.

References

Al Zamal, F.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Bakliwal, A.; Arora, P.; Madhappan, S.; Kapre, N.; Singh, M.; and Varma, V. 2012. Mining sentiments from tweets. *Proceedings of the WASSA* 12.

Bakliwal, A.; Foster, J.; van der Puil, J.; Orien, R.; Tounsi, L.; and Hughes, M. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. *NAACL 2013* 49.

Brown, P. F.; Desouza, P. V.; Mercer, R. L.; Pietra, V. J. D.; and Lai, J. C. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18(4):467–479.

Cohen, R., and Ruths, D. 2013. Classifying political orientation on twitter: It's not easy! In *ICWSM*.

Golbeck, J., and Hansen, D. 2011. Computing political preference among twitter followers. In *SIGCHI*, 1105–1108. ACM.

Miller, S.; Guinness, J.; and Zamanian, A. 2004. Name tagging with word clusters and discriminative training. In *HLT-NAACL*, volume 4, 337–342. Citeseer.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11:122–129.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *ACL EMNLP*, 79–86.

Pennebaker, J. W.; Chung, C. K.; Ireland, M.; Gonzales, A.; and Booth, R. J. 2007. The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net*.

Saraee, M., and Bagheri, A. 2013. Feature selection methods in persian sentiment analysis. In *Natural Language Processing and Information Systems*. Springer. 303–308.

Ushioda, A., and Kawasaki, J. 1996. Hierarchical clustering of words and application to nlp tasks. In *Proceedings of the Fourth Workshop on Very Large Corpora*, 28–41.

Wang, X.; Wei, F.; Liu, X.; Zhou, M.; and Zhang, M. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *CIKM*, 1031–1040. ACM.

Wang, H.; Can, D.; Kazemzadeh, A.; Bar, F.; and Narayanan, S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, 115–120.