Opinion Analysis of Bi-Lingual Event Data from Social Networks

Iqra Javed, Hammad Afzal

Department of Computer Software Engineering,
National University of Sciences and Technology, Islamabad, Pakistan
iqra217@gmail.com, hammad.afzal@mcs.edu.pk

Abstract. Social networks have recently emerged as the fastest and very effective medium to express news updates, trends and expression of personal views. There have been several studies to perform detailed sentiment analysis on such data in most of the developed languages. However, Urdu lacked any such study despite being spoken by around 30 Million people around the globe and used in regions with fastest growth of broadband users. This research has been carried out as a first step in this direction, where a language resource comprising the sentiment strengths of Roman Urdu words has been proposed along with its utility by under taking a case study of spatial analysis of bi-lingual (Urdu and English) tweets in the context of a national event, i.e. genral elections 2013. The results are encouraging, showing the effective utility of the bi-lingual sentiment strength database.

Keywords: Keywords: Sentiment Analysis, Twitter Data, Language Resources

1 Introduction

For last few years, there has been an emerging trend by public to consider the social networks for news updates, upcoming trends, community updates and expression of personal reviews on various events. These events range from smaller ones, interesting only to some particular region or community such as local seminars or concerts to the larger ones that can be of interest to entire country (epidemics, weather or political events). The popularity of social networks among public to share their opinion has led to its use as an opinion reviewing and result predicting tool for events that are related to public having common issues and problems. There have been several case studies that consider geographilcal and temporal analysis of such events [2-10]

Twitter[1] is considered as one of the most popular micro-blogging social networking website with more than 554 million active users till 2013[2]. Twitter user's posts, known as "tweets", are generally used as information broadcasting tool for local events and they can be used to mine their pre and post effects. In addition, they can also be used for opinion analysis from a specific region within specific time bounds.

[1] https://twitter.com/

[2] http://www.statisticbrain.com/twitter-statistics/

This research presents an approach on analysis of bi-lingual tweets, describing the public's opinions about a national event. We have particularly focused on a case study of Pakistan's general elections 2013. Pakistan has been considered as one of the fastest growing countries in terms of IT users and broadband usage. Youth being the major portion of population[3], such frameworks can be very effectively utilized for trend prediction. Although English is commonly used in higher education, public in general is not much well versed in English; however they are not restricted by this limitation and tend to express their opinions in Urdu using English script (termed as Roman Urdu hereafter in this paper). We have performed spatial and temporal analysis, covering five major cities in Pakistan (having populations around 50 Million each) and over the period of 5 months. The results obtained by our analyis mostly confirm with the results of elections (announced in March, 2013) and the observations made by other survey organizations (using the means other than social network data).

2 Background

Manually prepared lexicons and machine learning techniques have been mostly used in sentiment analysis to analyze mood, emotion classification and opinion extraction within a text provided tweets. In [2] proposed technique is based on classification of tweets on their content basis and groups them as hot topics according to the frequent population of tweets on relative topics and geo-location information associated with tweet text. However, due to semantic fluctuations, the proposed classification technique does not work particularly good enough as tweets can use multiple words to refer to the same event.

Ishikawa, Arakawa, Tagashira, Fukuda discusses a system that detects hot topic in a local area in a specified time period and a classification method is proposed that reduces variation of posted words related to the same topic in tweets. The hot topics can be predictable (matches, elections, festivals) and non-predictable (natural disasters) events. Such event analysis is helpful in making any business strategy, disease information social relationships [3].

Wong and Chang conducted quantitative and qualitative analysis on informative and affective tweets based on word frequencies and word co-occurrence [5]. They used event related context specific vocabulary to train their classifier. Open source resources have also been utilized for lexicon building and sentiment classification but the classifier gave poor performance on untrained domains [7]. Polarity classification was performed in [8] using lexicon-based approach where manual annotation was performed. They ruled out those tweets that contained both positive and negative emotions. Lexicon based approach is applied in Sentistrength [10] for sentiment analysis of text. But these lexicons provide limited support and needs manual marked lexicon. Further no support available for roman-Urdu and political text analysis.

[3] http://southasiainvestor.blogspot.com/2011/10/pakistan-ranks-among-fastest-growing.html

3 Methodology

The aim of the proposed research is to provide a framework to analyse the bi-lingual data from twitter using spatial and temporal bounds. Pakistan's general Election 2013 is taken as case study. Retrieved text from twitter comprises of tweets written in two languages, English and Roman-Urdu. The sentiment analysis is performed on this bi-lingual text using existing (customized) and newly created lexicons on sentiments data. The steps performed in our approach are illustrated in Fig 1 and elaborated below.

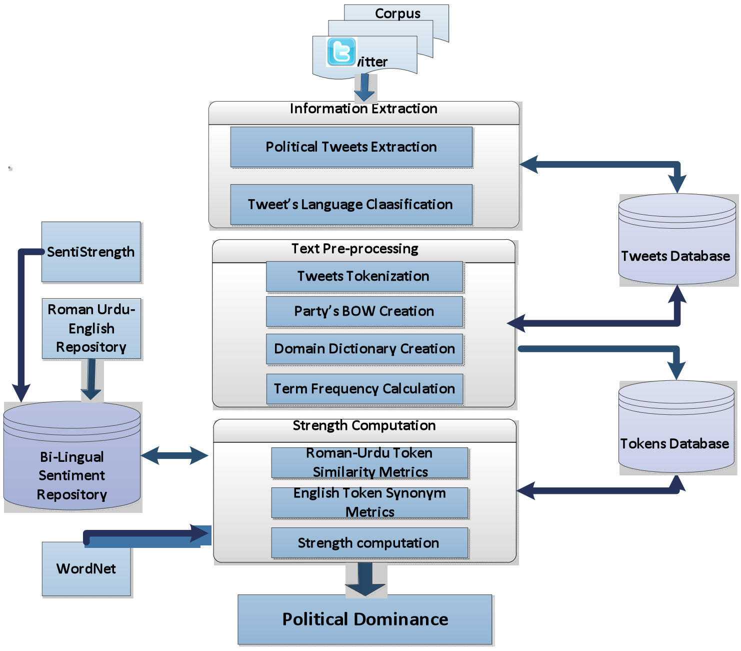
Corpus
/itter
Information Extraction
Political Tweets Extraction
Tweet's Language Claasification
Tweets Database
SentiStrength
Roman Urdu-English Repository
Text Pre-processing
Tweets Tokenization
Party's BOW Creation
Domain Dictionary Creation
Term Frequency Calculation
Tokens Database
Bi-Lingual Sentiment Repository
Strength Computation
Roman-Urdu Token Similarity Metrics
English Token Synonym Metrics
Strength computation
WordNet
Political Dominance

Fig. 1. Overview of Bi-lingual spatial-temporal event opinion analysis process

3.1 Collection of Bi-Lingual Tweets

Our approach starts with collection of tweets dataset. Twitter search API is used for tweets retrieval based on keywords. Tweets related to four main political parties Pakistan Tehreek-e-Insaaf (PTI), Pakistan Muslim League Nawaz PML(N), Pakistan Peoples Party (PPP) andMutahidda Quomi Movement (MQM) from five major cities of Pakistan (Islamabad, Lahore, Karachi, Peshawar and Quetta) considering the radius of 20 miles of the city are collected. Collection of dataset is performed on weekly basis while the time span for dataset collection is from Dec 2012 till polling day (11th March, 2013).

3.2 Classification of Tweets

Two iterations of classification are performed over dataset retrieved from twitter. These classifications are carried out on keyword basis. First iteration discriminates between the tweets belonging to political/non political contents. This step was required as most of the spammers, particularly belong to real estate businesses, exploited the popularity of the keywords related to political parties. Some keywords that were used to identify noisy (non political tweets) are summarized in Table 1.

Index	Noun	Verb
1.	bahria town	Sale
2.	Dha	Plot
3.	Villas	Buy
4.	Estate	Purchase
5.	Kanal	
6.	Marla	

Table 1. Keywords used to extract non-political tweets

Second iteration of classification was performed to discriminate between English and Roman-Urdu. This was also performed based on presence of keywords from a set of commonly used English words as presented in Table 2.

Adjectives	Adverbs	Conjunctions	Prepositions	Pronouns	Verbs
Good	Up	And	Of	It	Be
New	So	That	In	I	Have
First	Out	But	To	You	Do
Last	Just	Or	For	He	Say
Long	Now	As	With	They	Get

Table 2. Example of English Keywords Used For Language Classification.

S.No	Party	City	Language	Text
	Pti	Peshawar	Roman Urdu	peshawar: jamaat-e-islami aur pti ke dermian khyber pakhtunkhwa mey seat adjustment per ittefaak na husaka.
	Mqm	Karachi	Roman Urdu	karachi: mqm nay aam intikhabat main mulk bhar say party ticket kay liye darkhastain talab kar lein dr. farooq sattar.b.n
	Pml	Lahore	Roman Urdu	:lahore: \nsabiq governor state bank dr. ishrat hussain ko nigran wazir e azam banai janne ka imkaan zarai.\n#ppp #pmln #pti
	Pti	Islamabad	English	:#pti & #ji flirting in rawalpindi :d >>>> http:\/\/t.co\/0rqippguod

Table 3. Sample of Tweets Collected and Saved in Database.

3.3 Creation of Bi-Lingual Sentiment Repository

In order to perform text analysis of bi-lingual tweets, we need to develop a database that is capable of providing sentiment strength to words used within bi-lingual tweets messages. For English language, SentiStrength’[4] is used for extracting the English lexica’s sentiment strength. The original SentiStrength contains 2546 English words along with their sentiment score ranging from -4 to +4. However, there has not been any such attempt for Urdu (Roman Urdu) language. For this purpose, we created our own lexicon that provides the sentiment strength score to Roman Urdu words similar to the structure of SentiStrength. Two resources, SentiStrenght and English to Roman-Urdu dictionary[5] are utilized in order to create a unified sentiment strength database. English words from SentiStrength have been searched for their Roman-Urdu translations. English words with their Roman-Urdu translations are combined with SentiStrength to create **Bi-Lingual Sentiment Repository (BLSR)** as shown in Table 4.

Word	**Roman-Urdu Translations**			**Sentiment Strength**
	First	**Second**	**Third**	
Accident	Aafat	Haadisah	Ittefaaq	-2
Bury	dafan karna	Gaarna		-3
Callous	bey raehm	Sakht		-4
Calm	Aahistah	khaamosh		2
Delicious	Latiif	Laziiz	mazey daar	3
Excellent	Faazil	Khuub		4

Table 4. Example from Bi-Lingual Sentiment Repository (BLSR). Each English word is linked with three different Urdu translations (where available) along with the sentiment score.

Bi-Lingual Sentiment Repository (BLSR) thus created provides the sentiment strength of 1673 English as well as 3900 Roman-Urdu words. Sentiment strength ranges from -4 to -1 indicating negative strength (-4 as most negative and -1 as least negative) and 1 to 4 indicate positive strength(1 as least positive and 4 as most positive) where 0 represent no sentiment strength and behaves as neutral.

3.4 Sentiment Allocation and Computation

Tweets belonging to each political party are tokenized. After tokenization, each token is assigned strength from SentiStrength and BLSR. The strength of every single tweet is then computed as follows:

$$\text{Sentiment-Tweet (ST)} = \boldsymbol{F1 * S1 + F2 * S2 + F3 * S3 + \ldots\ldots\ Fn * Snn} \qquad (1)$$

[4] http://sentistrength.wlv.ac.uk/

[5] http://www.scribd.com/doc/14203656/English-to-Urdu-and-Roman-Urdu-Dictionary

Where,
F_1, F_2… F_n are the frequencies of the tokens appearing in a tweet,
S_1, S_2 … S_n are the sentiment strength of the corresponding token,
n is the number of tokens in a given tweet.

Using the database, the strength of each political party can then be computed as:

$$\text{Sentiment-Party (SP)} = \quad i=0mSTpim \qquad (2)$$

Where,
ST_{pi} is the strength of a tweet belonging to a particular party **p**.
m is the number of tweets belonging to party **p**.

3.5 Handling the Missing Tokens in BLSR

There are a lot of important terms that could not be found in BLSR because of typographical errors, transliteration errors as well as individual based short written English and Roman-Urdu words. To handle such typographical errors in Roman-Urdu tokens, a number of algorithms (Bigram-Based Cosine Similarity, Dice Coefficient and Jaccard Similarity) are applied for string approximation. We found that bigram-Cosine similarity outperformed other metrics.

To increase the recall of English words, WordNet is utilized to obtain synonyms for English tokens that did not exist in SentiStrength. Class sentiment strength is assigned to relevant tokens on the basis of synonyms.

4 Results and Discussion

The dataset contains 91,804 tweet messages collected for four political parties in five major cities along with noisy data (non-political) of 21,821 tweets. The detailed statistics regarding the number of tweets collected from various cities and about different parties is presented in Table 5.

Index	**City**	**Number of tweets collected**				**Total tweets**
		PTI	**PML**	**PPP**	**MQM**	
1	Islamabad	8534	3699	2606	2709	17548
2	Lahore	9903	7591	5719	7228	30441
3	Karachi	8763	2399	8572	7531	27265
4	Peshawar	9500	1755	2300	1476	15031
5	Queta	33	37	13	2	85

Table 5. Tweets Collection Statistics

In language classification 62797 tweets were classified as English and 7186 as Roman-Urdu tweet messages as depicted in Table 6.

Tweets category	Total
Total Dataset	91804
Political	69983
Non-Political	21821
English	62797
Roman-Urdu	7186

Table 6. Classification of tweet dataset

Table 7 represents the dominance of political parties in relevant cities based on sentiment analysis of roman-Urdu tweets. As described before, the results' coverage is improved by applying bigram-Cosine similarity metric on roman-Urdu tokens for removing typographical errors and similarity approximation. PTI is most dominant party in Queta and Islamabad whereas as PPP is most popular party in Peshawar using BLSR.

Index	City	Political Party dominance				No of tweets analyzed
		PTI	PML	PPP	MQM	
1	Islamabad	63%	2%	8%	27%	1291
2	Lahore	25%	10%	21%	46%	1936
3	Karachi	29%	14%	31%	26%	2921
4	Peshawar	3%	6%	97%	0%	587
5	Queta	100%	0%	0%	0%	40

Table 7. Political Dominance based on Sentiment strength analysis of Roman-Urdu Tweets

Table 8 depicts the dominance of political parties based on English tweets sentiment analysis using BLSR. PTI dominates other parties in general whereas in Lahore public opinion in mostly in favor of PML.

Index	City	Political Party dominance				No of tweets
		Pti	Pml	Ppp	Mqm	
1	Islamabad	62%	4%	3%	31%	4406
2	Lahore	5%	70%	10%	16%	6870
3	Karachi	38%	11%	19%	32%	8096
4	Peshawar	68%	14%	7%	11%	2276
5	Queta	23%	46%	30%	0%	23

Table 8. Political Dominance based on Sentiment strength analysis of English Tweets

5 Conclusions

We have proposed a method for sentiment analysis of bi-lingual, English and roman-Urdu data from social networks, particularly focusing on twitter data. We considered case study of general elections in Pakistan 2013. Tweets are collected related to major political parties of Pakistan considering four major cities. A bi-lingual lexicon is constructed that is capable of providing sentiment strength for English as well as roman-Urdu words used in tweets. In order to increase the coverage of this bilingual lexicon, WordNet is used to improve the performance of English tweets. Similarly, for Roman Urdu tweets, a bigram based consine similarity is used to reduce number of typographical errors as well as performing string approximation to increase the coverage. Using these resources, we have addressed the dominance of political parties in Pakistan before elections 2013. The difference in the results of English and Urdu Tweets shows the two separate clusters of population and their political affiliations. Furthermore, the inbalance between number of English and Urdu Tweets is because of simple classification method to detect language that has resulted in many Roman Urdu tweets marked as English. This could be improved by incorporating complex methodologies. Furthermore, the size of lexicon can be improved by using lexical and contextual similarity based techniques [11] to collect similar terms from a corpus (in this case, WWW can be used). The constructed bi-lingual lexicon is not domain specific and therefore, can be used for any other domain as well.

References

1. B. J. Jensen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth,"Journal of the American Society for Information Science and Technology, vol. 60, no. 11, pp. 2169–2188, 2009.
2. Chung-Hong Lee, Hsin-Chang, Tzan-Feng Chien and Wei-Shiang Wen Yang, "A Novel Approach for Event Detection by Mining Spatio-temporal Information on Microblogs," in International Conference on Advances in Social Networks Analysis and Mining, 2011.
3. Shota Ishikawa, Yutaka Arakawa, Shigeaki Tagashira, Akira Fukuda "Hot Topic Detection in Local Areas Using Twitter and Wikipedia," in ARCS Workshops (ARCS), 28-29 Feb. 2012.
4. Alexander Pak and Patrick Paroubek, "Twitter for Sentiment Analysis: When Language Resources Are Not Available," 22nd International Workshop on Database and Expert Systems Applications, 2011.
5. Yi Wu, Jackson Wong, Yimeng Deng, Klarissa Chang, "An Exploration of Social Media in Public Opinion Convergence: Elaboration Likelihood and Semantic Networks on Political Events," Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011.
6. Asli Celikyilmaz, Dilek Hakkani-Tur, Junlan Feng, "Probabilistic Model-Based Sentiment Analysis of Twitter Messages," Spoken Language Technology Workshop (SLT), 12-15 Dec. 2010:pp. 79 - 84.
7. Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, Jay Ramanathan, "Towards Building Large-Scale Distributed Systems for Twitter Sentiment Analysis," SAC'12, Riva del Garda, Italy, March 25-29, 2012,
8. Georgios Paltoglou and Mike Thelwall, "Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media," ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 4, Article 66, Publication date: September 2012.
9. Akshaya Iyengar, Tim Finin and Anupam Joshi, "Content-based prediction of temporal boundaries for events in Twitter," IEEE International Conference on Privacy, Security, Risk, Trust, and IEEE International Conference on Social Computing, 2011.

10. Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A. (2010).Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.

11. Hammad Afzal, Robert Stevens, Goran Nenadic: “Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary”, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine* (SMBM 2008): pp. 5-12