Advances in Digital Technologies
J. Mizera-Pietraszko et al. (Eds.)
IOS Press, 2016

doi:10.3233/978-1-61499-637-8-105

Bilingual Sentiment Detection – Investigating Impact of Tweet Translation

WANDEEP KAUR[a] and VIMALA BALAKRISHNAN[b]
[a,b] *University of Malaya, Malaysia*

Abstract. This paper looks at how translating tweets from Malay to English may impact its sentiment score, and if at all the score can be improved in case a tweet is translated compared to just translating the keywords. Tweets written in Malay were translated using an online dictionary before proceeding for analysis. An online sentiment analysis tool, Twinword was used to perform the sentiment analysis on both translated and untranslated tweets. The results of the analysis showed translating tweets did not create a significant impact on the overall sentiment score; therefore translating the whole length of the tweet would not affect the accuracy score.

Keywords. Sentiment detection, Twitter, bilingual tweets, tweet translation

1. Introduction

The internet today acts as a melting pot that encourages interactions to occur on a global scale regardless of one's geographical location. The boom of social media as well as ease of internet connection accessibility may it be via phone, tablet or laptop nurtures the privilege to voice out opinions, vent frustration and engage in online discussions across the border. The one element that allows such communication to take place is language. According to Internet World Stats, English is the predominant language used on the internet. Nevertheless, this does not discourage those who speak other languages to forge relationships and open communication channels with other language speaking counter parts over social media. For example, sport fans are able to discuss a particular match in more than one language ([1],[2]); in the marketing world, consumers are relying on the public feedback before purchasing a particular product or service ([3],[4]); people are also using the social media platform for disaster management ([5],[6]).

A survey conducted by Semiocast[1] (Social Media Intelligence Company) in 2013 revealed that the Malay language is fourth most widely language used on Twitter after English, Japanese and Spanish. The language is used widely within the Southeast Asian region covering Malaysia, Singapore, Indonesia and southernmost provinces of Thailand. However, there seem to be a lack of sentiment analysis done on tweets written in Malay. The main goal of this paper is to investigate the impact of translating a tweet from its original language to English.

[1] http://www.adweek.com/socialtimes/twitter-top-10-languages/494260

This paper is organized as follows. Section 2 highlights the literature reading which includes an overview of sentiment analysis as well as recent work in multiple language sentiment analysis domains. In section 3 the methodology used is introduced and the results are discussed in Section 4. Conclusion and future research is provided in Section 5.

2. Literature Review

2.1. Overview of Sentiment Analysis

According to [8], there are two approaches to sentiment analysis namely lexicon based approach and machine learning approach. However, sentiment analysis mainly relies on the machine learning approach; specifically the Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy (ME) approach ([9]). NB employs the properties of Bayes theorem and only needs a humble amount of training data to calculate its prediction parameters [8]. The SVM on the other hand is dependent on statistical learning concepts and it has the capacity to constitute a decision plane within the narrative feature by mapping data instances non-linearly to inner product space where classes can be detached directly with a hyperplane [8].

Sentiment analysis is defined as an area of research within the Natural Language Processing domain that emphasizes on mood identification or subjective elements within a text [7]. In general, sentiment classification can be categorized as positive, negative or neutral. The scale to which a classification can be done may vary (+1 to -1, +10 to -10) depending on the tools used. Sentiment analysis has repeatedly been used to display positive and negative trends in the data sets and has been accepted as a viable method to provide tentative insights into unstructured textual data ([9]).

2.2. Sentiment Analysis in multiple languages

The increase of internet connectivity from people of different ethnical and cultural background has opened the social media network to adapt to different languages used. Therefore, recent studies in the sentiment analysis domain have moved to include non-English tweets. A study done by [9] focussed on gathering opinions from multi-culture social media platforms and developing a bilingual method which would be able to mine Chinese and English tweets concurrently. Similarly, [10] proposed a polarity classification system that combines supervised and unsupervised learning using Spanish corpus of film reviews that were parallel translated into English. 10,000 random posts from nine different Facebook page in the Czech language were analysed and classified as positive, negative or neutral in another study by [11]. As the impact of globalization and internet accessibility reaches new horizon, it opens more doors to study its sentiment polarity individually. [7] published a paper on Arabic sentiment analysis. The research assigned scores to words found in the Arabic WordNet before performing a semi-supervised learning method to calculate accuracy of its sentiment scores.

3. Approach

Figure 1 shows the framework adopted for the proposed approach. The framework is further discussed in the following subsection.

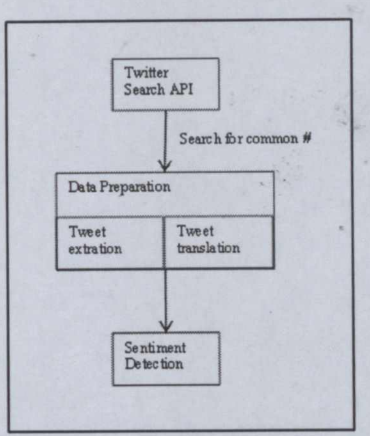

Figure 1. Framework for Sentiment Detection.

3.1. Data Preparation

A corpus of tweets was polled using the Twitter Search API during March and May 2015 for common hashtags related to the recent Government Service Tax (GST) implementation in Malaysia. The GST tax has only been enforced in Malaysia since April 2015. It's a relatively new tax introduction, one that has created debate among the general public. Hence, the sparked interest in collecting this data for analysis purpose. Table 1 shows the 9 common hashtags identified for this experiment. Approximately 5,200 tweets were extracted. Out of which tweets that were too short for analysis and tweets that contained more than two spelling errors were discarded. For the purpose of this experimentation, retweets were also discarded which were easily identified as those tweets began with RT. In the end, 3,600 tweets remained. From this sample, 1000 tweets were selected at random. Only tweets that were written in the Malay language were considered. This is solely for the investigation purpose of how translating a tweet from one language to another may impact on the overall sentiment score.

Table 1. Query terms used to poll tweets

#NoGST	#GSTMalaysia	#HapusGST
#BantahGST	#CabaranGST	#TolakGST
#MyGST	#GST	#WelcomeGST

3.2. Translating Tweets

The selected tweets were translated from Malay to English with the help of an expert. Each tweet was first translated manually using an online Malay dictionary, MalayCube[2]. Then the translated version of the tweets was validated to ensure the translated tweets did not lose its essence after translation. The sentiment of a tweet prior to translation was also noted. This was to create a check point of the tweet sentiment after translation.

3.3. Sentiment Detection

Once the tweets were translated, a free online sentiment analysis tool; Twinword[3] was used to detect the sentiment of the translated tweet. The tool categorizes the tweet into positive, negative or neutral based on the keywords extracted from a given text. If two or more words are positive then the sentiment returned is positive. Same goes to detecting a negative and neutral sentiment of a text whereby if two or more words are detected as negative, hence results would show the sentiment of the tweet is negative. Tweets that were translated from Malay to English were also validated for its sentiments using a human inter coder. Each tweet was read and categorized as positive or negative accordingly.

4. Results and discussion

Figure 2 shows the results of the sentiments after translation of 1000 random tweets selected. As per shown, the majority of the sentiment favored towards negative (497 tweets) compared to 310 positive tweets. Figure 3 on the other hand shows the percentage of tweets accordingly.

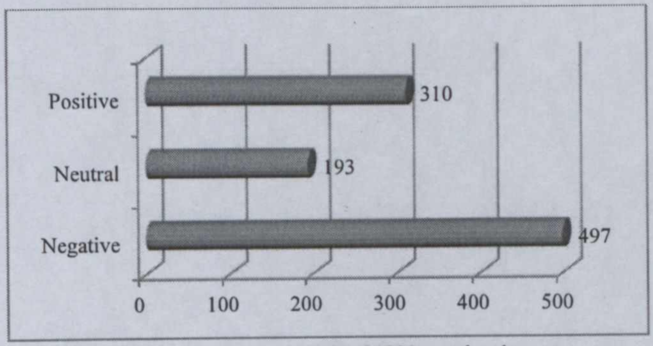

Figure 2. Sentiment analysis of 1000 translated tweets

[2] http://www.malaycube.com

[3] https://www.twinword.com/api/sentiment-analysis.php

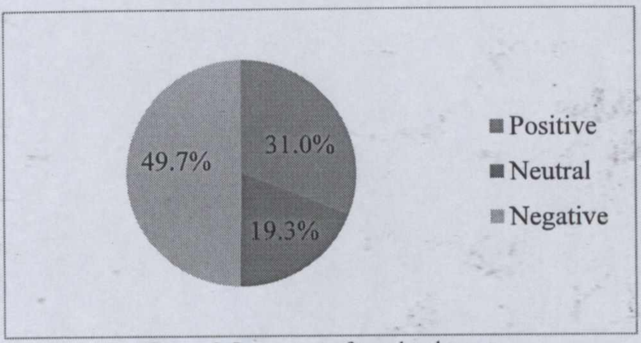

Figure 3. Percentage of translated tweets

This paper focuses on the impact of translation. Therefore in order to investigate if translation of tweets makes a difference, another analysis involving tweets translated from Malay to English was done which involved a human inter coder for the purpose of validation. The results of the human inter coder are presented as below.

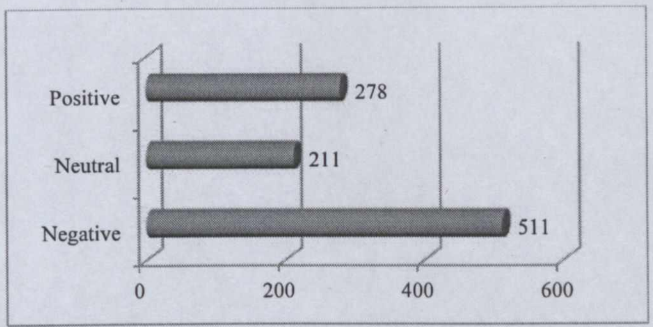

Figure 4. Sentiment analysis of 1000 translated tweets after human intervention

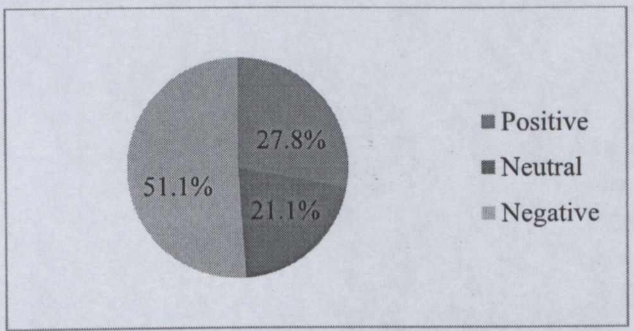

Figure 5. Percentage of translated tweets after human intervention

The result of the investigation showed that translation does not make a huge impact when it comes to sentiment analysis. Tweets that were annotated after translation differed about 4% compared to tweets that were directly translated and analyzed using

Twinword[4]. Similar experimentation was conducted on different corpus of tweets by [12] and [13]. Results obtained from the Chinese and Arabic tweets relatively were within 3% - 8% improvement when translated compared to untranslated analysis of tweet sentiment. Therefore, the method tested of translating the tweet as a whole does not significantly impact the overall sentiment analysis of a non-English corpus.

5. Conclusion

The purpose of this investigation was to study the significance of translating a tweet that is in a language other than English. Translating a word from non-English to English may cause it to lose its context as the same word may carry two different meanings depending on the sentence structure. Therefore this paper looked at the possibility of obtaining a different result if a sentiment analysis tool was fed with a tweet that was translated. The results discussed in this paper showed that the translation of a tweet does not considerably impact the overall results. The tool categorizes a tweet sentiment as positive, negative or neutral based on the words existing within its dictionary. If a word, may it be unigram, bigram or n-gram used within a tweet is classified as positive, the overall sentiment of the tweet would be returned as positive and so forth.

For future enhancement, the lexicon method can be applied to this corpus of data to compare how that method would impact the sentiment score. The analysis results of using a Malay dictionary compared to using an English dictionary tool with translated tweets can be compared to further validate the results of this investigation. Furthermore, a sentiment analysis using Malay lexicons could also be repeated to test the machine translation results when used to calculate the sentiment analysis percentage against tweets that were not translated as a whole.

6. Acknowledgement

The author wishes to extend their gratitude to University Malaya for supporting this study (RP028A-14AET).

7. References

[1] Corney, D., Martin, C., & Göker, A., Spot the ball: Detecting sports events on Twitter, *Advances in Information Retrieval* (2014), 449-454.

[2] Yu, Y., & Wang, X., World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets, *Computers in Human Behavior, 48*, (2015), 392-400.

[3] Chianasta, F., & Wijaya, S.,The impact of marketing promotion through social media on people's buying decision of Lenovo in internet era: A survey of social media users in Indonesia. *International Journal of Scientific and Research Publications, 4*(**1**), (2014), 1-6.

[4] Shin, H., Byun, C., & Lee, H., The Influence of Social Media: Twitter Usage Pattern during the 2014 Super Bowl Game. *International Journal of Multimedia & Ubiquitous Engineering, 10*(**3**), (2015)

[4] https://www.twinword.com/api/sentiment-analysis.php

[5] Chatfield, A. T., & Brajawidagda, U. *Twitter early tsunami warning system: A case study in Indonesia's Natural Disaster Management.* Paper presented at the System sciences (HICSS), 2013 46th Hawaii international conference on.

[6] Choi, S., & Bae, B., The Real-time Monitoring System of Social Big Data for Disaster Management *Computer Science and its Applications*, (2015), pp. 809-815

[7] Mahyoub, F. H., Siddiqui, M. A., & Dahab, M. Y., Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning. *Journal of King Saud University-Computer and Information Sciences,* ***26*****(4)**,(2014), pp. 417-424.

[8] Medhat, W., Hassan, A., & Korashy, H., Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal,* **5(4)**, (2014) 1093-1113.

[9] Yan, G., He, W., Shen, J., & Tang, C., A bilingual approach for conducting Chinese and English social media sentiment analysis. *Computer Networks, 75*, (2014), 491-503.

[10] Martín-Valdivia, M.-T., Martínez-Cámara, E., Perea-Ortega, J.-M., & Ureña-López, L. A., Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with applications,* ***40*****(10)**, (2013), 3934-3942.

[11] Habernal, I., Ptáček, T., & Steinberger, J., Supervised sentiment analysis in Czech social media. *Information Processing & Management,* ***50*****(5)**, (2014), 693-707.

[12] Lu, B., Tan, C., Cardie, C., & Tsou, B. K. *Joint bilingual sentiment classification with unlabeled parallel corpora.* (2011). Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

[13] Abdul-Mageed, M., Diab, M., & Kübler, S. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language,* ***28*****(1)**,(2014), 20-37.