A Review on Sentiment Analysis of Resource-Scarce Languages and Code-Mixed Social Media Text

Kathleen Swee Neo Tan[1] and Tong Ming Lim [2]

Department of Computer Science and Embedded Systems
Tunku Abdul Rahman University College
Kuala Lumpur, Malaysia
[1]tansn@tarc.edu.my, [2]limtm@tarc.edu.my

ABSTRACT. With the phenomenal increase of social media usage today, organizations and governments have more aggressively turned to analysing user-generated content freely available on social media platforms in order to gain insights and sentiments on products, brands and policies. One challenge faced by social media analytics researchers in this respect is the habit of users frequently expressing their opinions using code-mixed text in their posts and comments. Code-mixing refers to the mixing of words from multiple languages in a single sentence. The languages used may include both formal and informal languages, with at least one of the languages being a resource-scarce language. This paper presents a review on methods that have been proposed and used for sentiment analysis of code-mixed social media text.

KEYWORDS: code-mixed, resource-scarce, sentiment analysis, social media.

1 INTRODUCTION

Social media is becoming increasingly important in our daily lives. The number of social media users increased from 2.48 billion in 2017 to 2.65 billion in 2018 (www.statista.com). In addition, the number of monthly active Facebook users grew from 2.2 billion in 2018 to a staggering 2.45 billion as of third quarter 2019. This has led to the availability of a vast amount of user generated content (UGC), which has become a significant and free data source for companies, governments, and organizations to leverage on to obtain useful insights for decision-making.

Research in the area of sentiment analysis of social media text has predominantly been for text written in English (Balahur and Turchi, 2014). Recently, there has been a growing number of works with regards to sentiment analysis on other languages (Al-Saffar et al., 2018; Yue et al., 2019) as well as *code-mixed* text (Kaur and Mangat, 2017; Lo et al., 2016; Lo et al., 2017). *Code-mixed text* refers to text written in a mixture of languages. In social media, this often includes the use of informal languages such as slangs and misspelt words in the text.

This paper presents a review on the methods that have been proposed for the sentiment analysis of resource-scarce languages and code-mixed social media text. The rest of this paper is organized as follows: Section 2 provides a summary of methods that have been employed for sentiment analysis of text written in resource-scarce languages. Section 3 presents a review of the research that has been carried out on sentiment analysis of code-mixed text. Section 4 presents the conclusion.

2 SENTIMENT ANALYSIS OF RESOURCE-SCARCE LANGUAGES

2.1 Sentiment Analysis of Text in Resource-Scarce Languages

A major problem in performing sentiment analysis on languages other than English is the scarcity of sentiment resources (Araújoa et al., 2019; Vilares et al., 2017). Hence, one approach used by researchers is to first convert the text to

English via machine translation and then utilize English sentiment resources to determine the polarity of the text.

Balahur and Turchi (2014) carried out experiments with French, German and Spanish datasets that were translated to English using Bing Translator, Google Translate and Moses with unigrams, bigrams and term frequency-inverse document frequency (tf-idf) as features. However, the accuracy of the sentiment analysis is contingent on the level of accuracy provided by the translation tool.

Araújoa et al. (2019) carried out an extensive comparison of existing sentiment analysis tools for English as well as language-specific tools. They found that translating non-English text to English and then using sentiment analysis tools for the English language is a viable approach for languages that are supported by a proper machine translator.

Inaccurate results from machine translation are often used because some words do not have an equivalent word in English. This has spurred research for sentiment analysis in other languages that do not involve translation of the text to English beforehand. Bader et al. (2011) proposed an approach that does not require translation by using Latent Semantic Indexing (LSI) to project multilingual parallel corpus into a multilingual concept space. This enabled multilingual semantic comparisons to be carried out between documents and therefore removing the need for translation. The main disadvantage of this approach is that multilingual parallel corpus must be available, which may not be in the case for most languages and for social media text.

Boyd-Graber and Resnik (2010) proposed a multilingual supervised Latent Dirichlet Allocation (LDA) model that is able to connect the word distributions across German/English and German/Chinese language pairs. The advantage of their topic-modelling approach is that it does not require the use of machine translation nor parallel corpora. However, this approach ignores syntactic constructions and sentence structure, which may result in inaccurate sentiment determination. In addition, the findings indicate that the accuracy of prediction is very much influenced by the size of the training set.

A hybrid method combining Language Modelling (LM) implemented with a Logistic Regression classifier with a clue-based approach was proposed by Jeong et al. (2009). The LM approach comprised 162 different LMs that considered the various Korean syntactic categories, which is an indication of the need to comprehensively capture and represent the syntactic categories of the target languages for successful sentiment discovery. Although this method showed significant improvements over previous methods, it uses a very language-specific strategy and therefore, is not readily transferable to other languages.

Al-Saffar et al. (2018) proposed a hybrid approach which used a sentiment lexicon to extract features sets (sentiment words presence-level features, sentence-level features, sentiment words polarity level features, and subjective words conditional probability features) and a combination of supervised learning methods (Naïve Bayes, Support Vector Machine, and Deep Belief Network) with the final classification determined using majority voting. They found that the combination method together with various feature sets achieved the best result.

2.2 Building Lexicons for Resource-Scarce Languages

Sentiment analysis requires the use of sentiment resources to enable in order to determine the polarity of the text. As the manual creation of such resources is immensely expensive and time consuming, a number of researchers have looked into the problem of automatic construction of resources for sentiment analysis.

Methods to automatically build lexicons for any language would ideally exclude the use of machine translation and parallel corpora. To achieve this, the bootstrapping method using an initial small set of subjective words was employed by Volkova et al. (2013) to develop a

scalable and language independent model that is able to learn subjectivity clues from Twitter data. The results of their model's experiments on the Spanish and Russian languages showed that this approach could be further enhanced to generate sentiment lexicons for resource-scarce languages. Bakliwal et al. (2012) constructed a Hindi subjectivity lexicon using initial seed lists (positive, negative and objective words) and WordNet. Their approach performed reasonably well but needs to be further extended to incorporate morphological variations of words as well as word-sense disambiguation.

Xia et al. (2015) aimed to automatically construct versions of SenticNet (a concept-based sentiment analysis resource) for other languages. Experimenting with the Chinese language, they utilized online dictionaries and translation engines to translate English SenticNet to the target language before applying the use of topic modelling and sentiment orientation prediction algorithms for context disambiguation. Their work is promising as it incorporated word-sense disambiguation as well as language-dependent polarity and achieved acceptable results. The proposed model can be further extended by including automatic detection of language-dependent sentiment concepts.

3 SENTIMENT ANALYSIS ON CODE-MIXED SOCIAL MEDIA TEXT

In this section, a review of sentiment analysis of code-mixed social media text and key challenges encountered are presented.

3.1 Sentiment Analysis on Code-Mixed Text

Code-mixed text contains words from two or more languages in a single sentence. To perform sentiment analysis of Hindi-English text, the first phase of the method proposed by Sharma et al. (2016) consisted of performing language identification, slang detection, and word play correction using a dictionary-based approach. In the second phase, the sentiment analysis task was carried out using a lexicon-based approach which calculated the sentiments of the words using various sentiment resources. It was observed that inaccuracies in sentiment polarity classification occurred because the way that the sentences were constructed was not taken into consideration.

Kaur and Mangat's (2017) work proposed a dictionary-based technique which included the construction of a Hinglish (Hindi-English) dictionary using Wordnet, HindiWordnet, SentiWordnet and HindiSentiWordnet. Pravalika et al. (2017) built English, Hindi and slang dictionaries to serve as lexical resources for the sentiment analysis of Hindi-English code-mixed social media text. The dictionaries comprised English and Hindi words with their associated polarities which were integrated into a trie structure to ensure fast retrieval of words. Words for the Hindi and English slang dictionaries were extracted from the training corpus and merged into the trie. The incorporation of slangs, which include informal words and spelling variations, was necessary as their usage is rife on social media.

The work conducted by Lo et al. (2016) included the construction of a dictionary, a polarity lexicon as well as annotated dataset for Singlish (Singaporean English, which is essentially the combined use of English, Malay, Chinese and Chinese dialects) in a single sentence or message. They also proposed a hybrid system that includes the use of Singlish sentic patterns and Support Vector Machine (SVM). Their findings were that there is a need to consider code-mixing when analyzing localized languages.

Wang et al. (2017) considered Chinese-English text in their proposed joint factor graph model which aimed to identify the emotions expressed in the text. As they considered English as the embedded language, they used machine translation to translate individual English words to Chinese to form bigrams. Their experiment results showed that their proposed model achieved significant improvement in the emotions prediction of Chinese-English text.

The distributed representation of words in Hindi-English, Bengali-English and Kannada-English using Doc2Vec, FastText, Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) was investigated by Shalini et al.(2018). They identified the need to capture the semantics of code-mixed social media text in order to achieve better results in the task of sentiment classification.

Medrouk and Pappa (2018) investigated the use of deep neural networks (CNN and LSTM) on multilingual hotel and restaurant reviews written in English-Greek-French. Their work differed from existing methods as they used a unique multilingual and multi-domain corpus, and excluded the use of language modules. Their findings were that deep neural networks do not require complex language modules for sentiment polarity detection of code-mixed text.

Table 1: Sentiment analysis of code-mixed social media text

Reference	Language	Classification	Algorithms / Classifiers	Features Extracted	Lexical Resources	Dataset
Kaur & Mangat (2017)	Hindi-English (Hinglish)	Positive, negative	Dictionary-based	Unigram, bigram, trigram with Tf-idf	Wordnet, HindiWordnet, SentiWordnet, Hindi SentiWordnet	FB, Twitter, Youtube on movies
Lo et al. (2016)	English-Malay-Chinese-Chinese dialects (Singlish)	Positive, negative	Hybrid: SVM + Singlish sentic patterns	Unigram, bigram, trigram, emoticons, hybrid	Singlish annotated dataset	Twitter data with topics relevant to Singapore
Medrouk &Pappa (2018)	English-French-Greek	Positive, negative	CNN, LSTM	-	-	Restaurant & hotels reviews
Pravalika et al. (2017)	Hindi-English	Positive, negative, neutral	Hybrid: Lexicon-Based + Machine Learning (NB, SVM, Decision Tree, Random Tree, Multilayer Perceptron)	Unigram	Manually created English dictionary (words from Princeton), Hindi dictionary (IIT dictionary), and slangs from development corpus	FB posts on movies
Shalini et al. (2018)	Kannada-English, Bengali-English, Hindi-English	Positive, negative, neutral	Doc2Vec, Fasttext, CNN, Bi-LSTM	-	-	FB posts of news channels, SAIL 2017 code-mixed corpus
Sharma et al. (2016)	Hindi-English	Positive, negative, neutral	Dictionary-based, Lexicon-based	-	Opinion Lexicon, AFINN, WordNet, Hindi SentiWordNet	FIRE 2014, FIRE 2013, FB, Youtube
Wang et al. (2017)	Chinese-English	Happiness, sadness, anger, fear, surprise	Factor graph model + belief propagation	Bigram	MDBG CC-CEDICT bilingual lexicon, TongYiCiLin Chinese synonym dictionary	Weibo posts

Tf-idf – Term Frequency-Inverse Document Frequency
FIRE – Forum for IR Evaluation
SAIL – Sentiment Analysis for Indian Languages (SAIL)

At present, there is limited work on sentiment analysis of code-mixed social media text. In addition, the majority of work was focused on determining the sentiment polarity (i.e. positive, negative or neutral) of text and there is very little research on fine-grained sentiment analysis that is concerned with emotion classification (e.g., happiness, sadness or anger). Table 1 provides a summary of related work in this area so far.

3.2 Key Challenges

The first key challenge encountered in the sentiment analysis of code-mixed text is the scarcity of sentiment analysis resources. In addition, given the extensive use of informal languages or slangs on social media, there is also a need for lexical and sentiment resources for informal languages to enable the sentiments being expressed in the written text to be

determined. To avoid manual constructions of such resources, the automation of sentiment lexicons is a definite necessity.

The second challenge faced in the sentiment analysis task for code-mixed text is that a word may appear in multiple languages. For example, the word "*air*" in English means "*water*" in the Malay language. Thus, there is a need to carry out *language disambiguation* in addition to word-sense disambiguation.

Thirdly, the same word may invoke different sentiments in different languages. For example, the word "*dragon*" would have a negative sentiment in the English language but a positive sentiment in the Chinese language. Hence, this poses the need for language-specific sentiment concepts.

Traditionally, supervised learning methods are able to provide higher accuracy for the sentiment analysis task compared to unsupervised learning methods. Their reliance on labelled corpora, however, is often a hindrance to their use as there is limited or even no such existence of labelled datasets for the code-mixed text combining the use of specific languages. Thus, supervised learning methods may not be feasible as it is extremely expensive and time-consuming to produce manually annotated corpora. In view of this, semi-supervised and unsupervised learning methods, particularly in the use of neural networks and distributed word vector representations, could be explored.

4 CONCLUSION

As the Internet penetration rate continues to grow at an unprecedented rate around the world and the numbers of social media users in each country rise in tandem, stakeholders in each nation seek to capitalize on the capabilities of sentiment analysis systems to provide insights about the sentiments of customers or the general public on various matters. This paper provides a review of methods employed for the sentiment analysis on resource-scarce languages and code-mixed social media text. In addition, the challenges discussed in this paper forms future research opportunities and gaps to be addressed in the area of code-mixed sentiment analysis research undertakings.

Future directions would be to look into the automatic construction of sentiment resources for resource-scarce and code-mixed languages (which includes slang) using semi-supervised or unsupervised machine learning techniques in order to reduce dependence on manually created sentiment resources. The challenges with regards to language disambiguation, word-sense disambiguation, and language-specific sentiment concepts also need to be addressed. In addition, given that it may not be feasible to have available labelled code-mixed datasets for training and the challenges of constructing language models for informal languages, the use of deep neural networks may be explored for solving these problems.

REFERENCES

Al-Saffar, A. et al., 2018. Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLoS ONE*, 13(4), pp.1–18.

Araújoa, M., Pereirab, A. and Benevenuto, F., 2019. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, (xxxx).

Bader, B.W., Kegelmeyer, W.P. and Chew, P.A., 2011. Multilingual sentiment analysis using Latent Semantic Indexing and machine learning. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp.45–52.

Bakliwal, A., Arora, P. and Varma, V., 2012. Hindi subjective lexicon: A lexical resource for Hindi polarity classification. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, (May), pp.1189–1196.

Balahur, A. and Turchi, M., 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1), pp.56–75. Available at:

http://dx.doi.org/10.1016/j.csl.2013.03.004.

Boyd-Graber, J. and Resnik, P., 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. *EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (October), pp.45–55.

Jeong, Y. et al., 2009. Generating and mixing feature sets from language models for sentiment classification. *2009 International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2009*, pp.1–8.

Kaur, H. and Mangat, V., 2017. Dictionary based Sentiment Analysis of Hinglish text. *International Journal of Advanced Research in Computer Science*, 8(5), pp.816–822. Available at: www.ijarcs.info.

Lo, S.L., Cambria, E., Chiong, R. and Cornforth, D., 2016. A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection. *Knowledge-Based Systems*, 105, pp.236–247.

Lo, S.L., Cambria, E., Chiong, R. and Cornforth, D., 2017. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4), pp.499–527.

Medrouk, L. and Pappa, A., 2018. Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification? *Proceedings of the International Joint Conference on Neural Networks*, 2018-July, pp.1–6.

Pravalika, A., Oza, V., Meghana, N.P. and Kamath, S.S., 2017. Domain-specific sentiment analysis approaches for code-mixed social network data. *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*.

Shalini, K., Ganesh, H.B.B., Kumar, M.A. and Soman, K.P., 2018. Sentiment Analysis for Code-Mixed Indian Social Media Text with Distributed Representation. *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, pp.1126–1131.

Sharma, S., Srinivas, P.Y.K.L. and Balabantaray, R.C., 2016. Sentiment analysis of code - Mix script. *2015 International Conference on Computing and Network Communications, CoCoNet 2015*, pp.530–534.

Vilares, D., Alonso, M.A. and Gómez-Rodríguez, C., 2017. Supervised sentiment analysis in multilingual environments. *Information Processing and Management*, 53(3), pp.595–607. Available at: http://dx.doi.org/10.1016/j.ipm.2017.01.004.

Volkova, S., Wilson, T. and Yarowsky, D., 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2, pp.505–510.

Wang, Z., Lee, S.Y.M., Li, S. and Zhou, G., 2017. Emotion Analysis in Code-Switching Text with Joint Factor Graph Model. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(3), pp.469–480.

Xia, Y., Li, X., Cambria, E. and Hussain, A., 2015. A localization toolkit for sentic net. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2015-Janua(January), pp.403–408.

Yue, L. et al., 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2), pp.617–663. Available at: https://doi.org/10.1007/s10115-018-1236-4.