Proceedings of the LREC 2020 – 4th Workshop on Computational Approaches to Code Switching, pages 45–51
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
European Language Resources Association (ELRA), licensed under CC-BY-NC

Sentiment Analysis for Hinglish Code-mixed Tweets by means of Cross-lingual Word Embeddings

Pranaydeep Singh and Els Lefever
LT3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
pranaydeeps@gmail.com, els.lefever@ugent.be

Abstract

This paper investigates the use of unsupervised cross-lingual embeddings for solving the problem of code-mixed social media text understanding. We specifically investigate the use of these embeddings for a sentiment analysis task for Hinglish Tweets, viz. English combined with (transliterated) Hindi. In a first step, baseline models, initialized with monolingual embeddings obtained from large collections of tweets in English and code-mixed Hinglish, were trained. In a second step, two systems using cross-lingual embeddings were researched, being (1) a supervised classifier and (2) a transfer learning approach trained on English sentiment data and evaluated on code-mixed data. We demonstrate that incorporating cross-lingual embeddings improves the results (F1-score of *0.635* versus a monolingual baseline of *0.616*), without any parallel data required to train the cross-lingual embeddings. In addition, the results show that the cross-lingual embeddings not only improve the results in a fully supervised setting, but they can also be used as a base for distant supervision, by training a sentiment model in one of the source languages and evaluating on the other language projected in the same space. The transfer learning experiments result in an F1-score of *0.556* which is almost on par with the supervised settings and speak to the robustness of the cross-lingual embeddings approach.

Keywords: sentiment analysis, code-mixed text, Hinglish, cross-lingual word embeddings, transfer learning

1. Introduction

Code-mixing is a frequent phenomenon in user-generated content on social media. In linguistics, code-mixing traditionally refers to the embedding of linguistic units (phrases, words, morphemes) into an utterance of another language (Myers-Scotton, 1993). In that sense, it can be distinguished from *code-switching*, which refers to a “juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems” (Gumperz, 1982), where the alternation usually takes the form of two subsequent sentences. In the proposed research, code-mixing is considered as a phenomenon where linguistic units in Hindi are embedded in English text, or the other way around, but this can take place both at the sentence and word level. As a consequence, we will use the term *code-mixing* as an umbrella term that can imply both linguistic phenomena.

The phenomenon of code-mixing frequently occurs in spoken languages, such as for instance a combination of English with Spanish (so-called *Spanglish*) or English with Hindi (so-called *Hinglish*). More recently, due to the rise of the web 2.0 and the proliferation of user-generated content on the internet, it is increasingly used in written text as well. This social media content is very important to automatically analyse the public opinion on products, politics or events (task of sentiment analysis), to analyse the different emotions of the public triggered by events (task of emotion detection), to observe trends, etc. Code-mixing is, however, very challenging for standard NLP pipelines, which are usually trained on large monolingual resources (e.g. English or Hindi). As a result, these tools cannot cope with code-mixing in the data. In addition, social media language is characterized by informal language use, containing a lot of abbreviations, spelling mistakes, flooding, emojis, emoticons and wrong grammatical constructions. In the case of Hinglish, an additional challenge is added because people do not only switch between languages (e.g. English and Hindi), but also use English phonetic typing to write Hindi words, instead of using the Devanagari script.

In this paper, we propose a sentiment analysis approach for Hinglish tweets, containing a mix of English and transliterated Hindi. To this end, cross-lingual word embeddings for English and transliterated Hindi are constructed. The proposed research has been carried out in preparation of experiments for the SemEval 2020 shared task on sentiment analysis in code-mixed social media text (Das et al., 2020). This task consists of predicting the sentiment (positive, negative, neutral) of a given code-mixed tweet. Whereas the SemEval task is designed for both English-Hindi and English-Spanish, we will only investigate sentiment analysis for English-Hindi code-mixed tweets in this research.

The remainder of this paper is organized as follows. In Section 2., we summarize relevant related research, whereas Section 3. gives an overview of the data set used to train and evaluate the system. Section 4. describes our approach to sentiment analysis for code-mixed Hinglish data. In section 5., we report on the results and provide an analysis of the performance, while Section 6. concludes this paper and gives directions for future research.

2. Related Research

Related research on computational models for code-mixing is scarce because of the rarity of the phenomenon in conventional text corpora, which makes it hard to apply data-greedy approaches. Previous research, however, has tried to predict code-switching in English-Spanish (Solorio and Liu, 2008a; Solorio and Liu, 2008b) and Turkish-Dutch (Nguyen and Seza Dogruoz, 2013) text corpora.

More recently, research has been performed to study code-switching on social media from a computational angle. Vyas et al. (2014) have compiled an annotated corpus for Hindi-English from Facebook forums, and performed experiments for language identification, back-transliteration, normalization and part-of-speech tagging on this corpus. They identify normalisation and transliteration as very challenging problems for Hinglish. Similar work has been carried out by Sharma et al. (2016), who developed a shallow parser for Hindi-English code-mixed social media text. Rijhwani et al. (2017) introduce an unsupervised word-level language detection technique (using a Hidden Markov Model) for code-switched text on Twitter that can be applied to different languages.

Pratapa et al. (2018) compare three bilingual word embedding approaches, bilingual correlation based embeddings (Faruqui and Dyer, 2014), bilingual compositional model (Hermann and Blunsom, 2014) and bilingual Skip-gram (Luong et al., 2015), to perform code-mixed sentiment analysis and Part-of-Speech tagging. In addition, they also train skip gram embeddings on synthetic code-mixed text. Their results show that the applied bilingual embeddings do not perform well, and that multilingual embeddings might be a better solution to process code-mixed text. This is mainly due to the fact that code-mixed text contains particular semantic and syntactic structures that do not occur in the respective monolingual corpora.

Seminal work in sentiment analysis (SA) of Hindi text was done by Joshi et al. (Joshi et al., 2010), who built a system containing a classification, machine translation and sentiment lexicon module. Bakliwal et al. (2012) created a sentiment lexicon for Hindi, and Das and Bandyophadhyay (2010) created the Hindi SentiWordNet.

Joshi et al. (2016) introduce a Hindi-English code-mixed dataset for sentiment analysis and propose a system to SA that learns sub-word level representations in LSTM (Long Short-Term Memory) (Subword-LSTM) instead of character- or word-level representations.

Due to the unavailability of NLP tools for Hinglish code-mixed data, we cannot apply a standard sentiment analysis pipeline. To overcome this, we propose a novel method to SA for Hinglish code-mixed tweets that applies cross-lingual word embeddings. To this end, we train monolingual embeddings for code-mixed data using independently gathered Twitter data, and then align the said monolingual embeddings with pre-trained English embeddings. This enables our models to learn from the encapsulated knowledge in pre-trained English embeddings without having much information about the code-mixed structure. Not only does this allow us to build a system that can perform sentiment analysis on bilingual data, but it also enables us to build a transfer learning based system that can derive information from a model trained in one language, to perform predictions in another language.

Most past work building cross-lingual sentiment models does so using translation systems (Zhou et al., 2016) or cross-lingual signals in another form, such as parallel corpora or bilingual dictionaries (Chen et al., 2018). However, since we work with code-mixed (transliterated) Hinglish Twitter data, there are no available resources like parallel corpora or bilingual dictionaries. Moreover, the ever evolving nature of social media text and various spelling alternatives in code-mixed data would make data greedy approaches like parallel corpora redundant.

In the proposed research, we thus build upon the recent research in constructing unsupervised cross-lingual embeddings by exploiting the inherent spacial structural similarity of word embeddings. Mulitple approaches use adversarial learning to learn these mappings with different ideas for optimization. While Zhang et al. (2017) choose to use Earth Mover's Distance as a similarity metric between two embedding spaces, Conneau et al. (2017) opt for the Procrustes solution to refine the mappings. In our experiments, we compare the results obtained when applying (1) the approach of Artexte et al. (2018), which uses Singular Value Decomposition and synthetic bilingual dictionary induction using similarity distributions, and (2) the approach of Conneau et al. (2017). We demonstrate that aligning code-mixed social media text with an anchor language like English helps to increase the performance in both a supervised and transfer learning setting.

3. Data

To train and evaluate our sentiment analysis system for Hinglish, we use the training data provided for the SemEval 2020 shared task on sentiment analysis in code-mixed social media text (Das et al., 2020). This dataset for Hinglish contains 15,131 instances, which have been labeled as positive, negative, or neutral. Besides the sentiment labels, the organisers also provide the language labels at the word level, consisting of the following tags: *en* (English), *hi* (Hindi), *mixed* and *univ* (e.g., symbols, @ mentions, hashtags). Table 1 shows some examples of the Hinglish code-mixed data, whereas Table 2 lists the statistics of the data set used for the sentiment analysis experiments.

As mentioned before, the data set contains a mixture of English and romanized or transliterated Hindi. This produces an additional challenge, as this romanized code-mixed data contains non-standard spellings like *aapke* and *apke* ("your"), non-grammatical constructions like *"Wow the amusement never ends even after the election Daily soap bana ke rakh diya"* which combines an English sentence with a Hindi sentence mid-way, and words which combine an English word with a Hindi alteration like *Jungli* ("wild") and *Filmy* ("glamorous"). Although the data set is tagged with a language label for every word, we did not use this information in our experiments as our aim was to build a common bilingual model that would be applicable for other code-mixed data sets as well.

4. Sentiment Analysis for Hinglish

This research aims to investigate the effectiveness of cross-lingual embeddings to perform sentiment analysis for code-

Tweet	@	Atheist	_	Krishna	JCB	full	trend	me	chal	rahi	hai	Label
Language Tag	Univ	En	Univ	En	En	En	En	En	Hi	Hi	Hi	Positive

Tweet	@	tamashbeen	_	Well	chara	Chor	ke	chele	this	news	is	a	year	old	...	Label
Language tag	Univ	En	Univ	En	Hi	Hi	Hi	En	En	En	En	En	En	En	Univ	Negative

Tweet	@	ur	_	boi	_	kdo	Most	unpractical	and	cool	sword	I	’	ve	seen	Label
Language Tag	Univ	Hi	Univ	Hi	Univ	Hi	En	En	En	En	En	En	Uni	En	En	Neutral

Table 1: Some Examples from the SemEval 2020 Code-Mixed Hinglish Challenge Dataset

Language Labels	
English Words	27,594
Hindi Words	28,167
Universal Symbols	2,792

Sentiment Labels	
Positive Tweets	5,034
Negative Tweets	4,459
Neutral Tweets	5,683

Table 2: Overview of the statistics of the data set used to perform Hinglish code-mixed sentiment analysis.

mixed data. Since the objective is to demonstrate the viability of cross-lingual embeddings over the simpler, monolingual embeddings, the experimental protocol dictates that the same classifier must be used to evaluate the systems. For the purpose of classification, we opted to use a Bi-LSTM encoder followed by a Softmax layer. Pre-trained crosslingual or monolingual embeddings were fed to the LSTM, the size of the hidden layer was 128 and we incorporated 4 layers in our model. This was followed by a single linear layer and the whole system was trained with Cross-Entropy Loss optimized with Stochastic Gradient Descent (SGD). Each of the models was trained and evaluated with 5-fold cross-validation, and an internal 5-fold cross-validation was performed on the training partition for hyper-parameter optimization.

We investigated two different methods to train our sentiment analysis system for Hinglish code-mixed tweets and compared them with monolingual baseline systems, resulting in the following three experimental setups:

1. Baseline Monolingual Systems: Models exclusively trained using monolingual embeddings
2. Supervised Classification: Models incorporating cross-lingual English-transliterated Hindi embeddings
3. Transfer Learning: Models trained with no supervision on the Hinglish data set but deriving knowledge from the English sentiment data sets

4.1. Baseline Systems with Monolingual Embeddings

Our baseline models were trained with monolingual embeddings in both languages, viz. code-mixed Hindi (Baseline H) and English (Baseline E). To train these monolingual embeddings, we first scraped tweets by means of the Twitter API in both English and transliterated Hindi. For English 141,566 tweets were scraped, while 252,183 tweets were scraped for Hindi. Hinglish tweets were obtained from the API by querying Hindi tweets and then filtering out tweets containing any Devanagari characters. We were left with 138,589 tweets for Hinglish after removing these ‘Devanagari’ tweets. Subsequently, monolingual embeddings were trained for both of the above mentioned corpora with a continuous bag-of-words FastText model (Bojanowski et al., 2017), and used to train a bi-directional LSTM (as explained above).

4.2. Supervised Sentiment Analysis with Cross-lingual Embeddings

Cross-lingual embeddings rely on the inherent similarities in language structure and composition to project multiple monolingual embeddings into the same space, enabling tasks which require knowledge of more than one language (Conneau et al., 2018). This kind of embeddings have been used to solve a variety of tasks like word-to-word translation (Chen and Cardie, 2018), evaluating sentence similarity (Bjerva and Östling, 2017) and detecting cognates across languages (Labat and Lefever, 2019). Most methods to project two or more monolingual embeddings into a shared space require a parallel seed dictionary to initialize an alignment which can then be improved upon (Upadhyay et al., 2016). The latter approach is not feasible, though, in this particular setting, as we aim to align English words with code-mixed (transliterated) Hinglish words, which often have no standardised spelling, but on the contrary occur with many variations in social media data. In recent research, however, a number of methods have been explored that seek to create a projection without any seed dictionary by relying on certain basic characteristics of a language in an embedding space. For our experiments, we evaluated two of these methods, namely the Multilingual Unsupervised and Supervised Embeddings (MUSE) Python library[1] and the VecMap toolkit[2], to create cross-lingual embeddings. We selected these methods in particular because of high performance in a number of downstream cross-lingual tasks and the lack of parallel data required to train the cross-lingual embedddings.

The **MUSE** ((Multilingual Unsupervised and Supervised Embeddings) toolkit (Lample and Conneau, 2019) uses a domain-adversarial setting to compensate for the lack of supervision. If the mapping matrix is referred to as W, and the

[1] https://ai.facebook.com/tools/muse/

[2] https://github.com/artetxem/vecmap

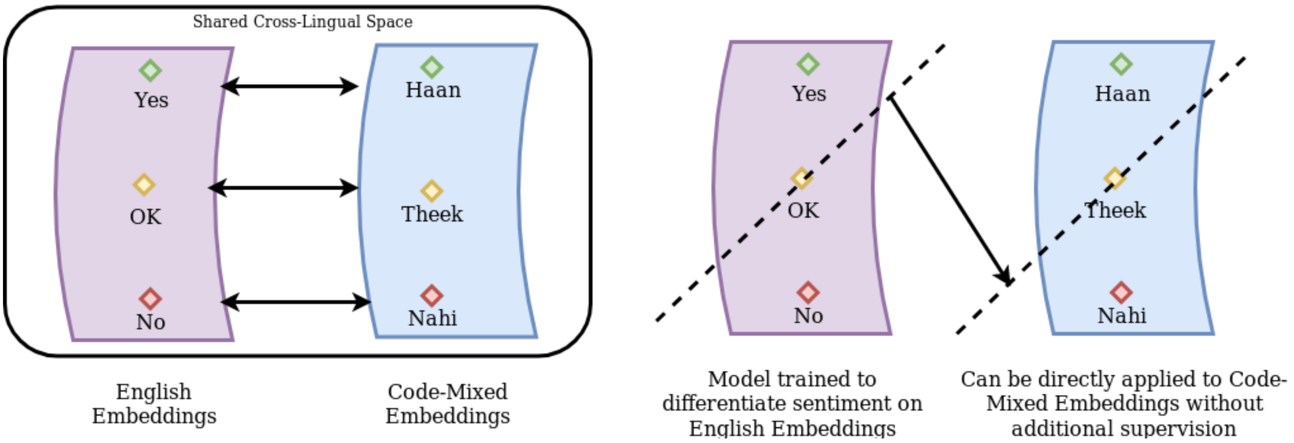
Shared Cross-Lingual Space
Yes
Haan
OK
Theek
No
Nahi
English Embeddings
Code-Mixed Embeddings
Yes
OK
No
Haan
Theek
Nahi
Model trained to differentiate sentiment on English Embeddings
Can be directly applied to Code-Mixed Embeddings without additional supervision

Figure 1: Transfer Learning based Sentiment Analysis for Hinglish, using cross-lingual embeddings

respective monolingual embeddings are referred to as X and Y, then the discriminator is trained to distinguish between WX and Y, whereas W is trained to prevent the discriminator from making accurate predictions by aligning WX and Y as closely as possible. Moreover, an iterative refinement tool using the Procrustes solution is used to further improve the alignment using synthetic dictionaries created from the most frequent words.

The **VecMap** toolkit (Artetxe et al., 2018), on the other hand, starts from the principle that if a similarity matrix of all words in a vocabulary was to be created, then every word would have a unique distribution and that this distribution would be consistent across languages. This principle is used to induct an initial seed dictionary. Optimal orthogonal mappings are then computed using Singular Value Decomposition while iteratively using the improved seed dictionary created by the current mapping. Multiple tweaks to the method, like bi-directional induction of the seed dictionary and symmetric re-weighting of the target language embeddings according to cross-correlation, further improve the quality of the mappings.

For our experiments, we tested two variants of both the VecMap and MUSE cross-lingual embeddings: (1) embeddings aligned with an entirely unsupervised dictionary induction method and (2) embeddings aligned using numerals and common tokens like “https” as a bilingual seed dictionary. This methods is especially interesting to look at as there is a decent overlap between the vocabulary of both embeddings as Hinglish is a derivative of English. The classifiers were then trained and tested by means of 5-fold cross-validation on the SemEval 2020 data.

4.3. Transfer Learning with Cross-lingual Embeddings

Approaches like VecMap and MUSE allow us to find an alignment which transforms monolingual embedddings into a shared space. Since this projection is done with no supervision (or minimal supervision in the case where numerals and identifiers are used as a seed dictionary), it should also be possible to train sentiment models for one of the languages and evaluate them on the other language. This can work if we assume that the model learns the sentiment-related information in the shared space in which both languages reside. To test these assumptions, we train a bi-directional LSTM on the English sentiment data of the SemEval-2016 “Sentiment Analysis in Twitter” task (Nakov et al., 2016) using English embeddings in the same shared space as code-mixed Hinglish embeddings. We then evaluate the model on the SemEval-2020 Hinglish data set, using the Hinglish embeddings pre-aligned with English embeddings.

Figure 1 illustrates the intuition behind this experiment. Since the model learns to associate particular words to particular sentiments in English during the supervision step, it should ideally also pick up the corresponding words and their sentiments in the code-mixed data due to the shared space, and by consequence be able to perform sentiment analysis with no direct supervision in the code-mixed data. As in the supervised setting (see Section 4.2.), we test for embeddings aligned with VecMap and MUSE, using both (1) the completely unsupervised (*Unsupervised*) and (2) the numerals and special characters seed methods (*SeedDict*).

5. Classification Results

Table 3 gives an overview of the results for supervised sentiment analysis when incorporating monolingual and various flavours of cross-lingual embeddings, while Table 4 shows the results when training a sentiment analysis system on English data (SemEval-2016) and applying it on the code-mixed data set (SemEval-2020). The experimental results for both system architectures reveal a number of interesting outcomes.

Firstly, it can be noted that the SeedDict VecMap approach consistently outperforms other types of cross-lingual embeddings. While for the supervised experiments, the cross-lingual embeddings do not outperform classical embeddings by a large margin, there are small improvements which can be accounted for by the fact that we can use both English as well as code-mixed embeddings to classify a sentence, whereas only one of those can be used at a time in standard monolingual approaches. While the quality of the embeddings may have diminished due to the alignement process, the results are still better due to the increased vocabulary at our disposal.

A tweet like *“One India sabka saath sabka vikas sabka*

	Positive			Negative			Neutral			Macro-Average		
Experiment	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
English (Baseline E)	0.734	0.574	0.645	0.625	0.633	0.629	0.501	0.595	0.544	0.620	0.600	0.606
Code-Mixed (Baseline H)	0.719	0.620	0.666	0.627	0.656	0.641	0.521	0.566	0.543	0.622	0.614	0.616
MUSE Unsupervised	0.750	0.539	0.627	0.612	0.735	**0.668**	0.511	0.557	0.533	0.624	0.610	0.609
MUSE SeedDict	**0.759**	0.540	0.631	**0.732**	0.528	0.614	0.500	**0.744**	**0.598**	**0.663**	0.604	0.614
VecMap Unsupervised	0.693	**0.691**	0.692	0.570	**0.804**	0.667	**0.565**	0.378	0.453	0.609	0.624	0.604
VecMap SeedDict	0.702	0.684	**0.693**	0.669	0.622	0.645	0.546	0.590	0.567	0.639	**0.632**	**0.635**

Table 3: Precision (Prec), Recall (Rec) and F1-score for all three sentiment classes for the Bidirectional LSTM models trained with various embedding flavours incorporated in a **supervised** system architecture for sentiment analysis for Hinglish.

	Positive			Negative			Neutral			Macro-Average		
Experiment	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score	Prec	Rec	F-score
MUSE Unsupervised	0.570	0.577	0.573	0.523	0.670	0.588	0.428	0.327	0.371	0.507	0.524	0.510
MUSE SeedDict	0.603	0.621	0.612	0.507	**0.789**	0.618	0.449	0.239	0.312	0.519	0.549	0.514
VecMap Unsupervised	0.580	**0.716**	**0.641**	**0.548**	0.688	0.610	0.457	0.268	0.338	0.528	0.557	0.529
VecMap SeedDict	**0.688**	0.529	0.598	0.541	0.748	**0.628**	**0.469**	**0.423**	**0.444**	**0.566**	**0.566**	**0.556**

Table 4: Precision (Prec), Recall (Rec) and F1-score for all three sentiment classes for the **transfer learning** sentiment systems trained on the SemEval-2016 English Twitter Data and evaluated on the SemEval-2020 code-mixed Hinglish Data.

visvas"(One India, with togetherness, progress and trust) is misclassified by the model incorporating monolingual english embeddings as "Neutral" since it cannot pick up the positive code-mixed Hindi words, while a tweet like *"FF Have a great weekend"* is misclassified by the monolingual code-mixed embeddings because of lack of knowledge of English words. Both of these tweets are, however, correctly classified by the VecMap embeddings using a seed dictionary.

It can also be observed that our transfer learning based model is able to perform sentiment analysis with acceptable accuracies without needing code-mixed supervision of any degree. This is a very promising outcome for low(er)-resourced languages, where large dedicated data sets for NLP tasks such as sentiment analysis are lacking. Regarding the baseline approaches, it is also worth noting that the Code-Mixed Baseline does not perform a lot better than the English baseline as one would expect. This can probably be attributed to the quality of the monolingual embeddings, since the English embeddings were trained on the vast Common Crawl data while the Code-Mixed embeddings were trained on a little more than 100,000 scraped tweets. While the classification is understandably accurate for tweets containing a majority of English words like *"Exclusive censor reports of Bharat is world class Words like movie of the year"* and less reliable for sentences predominantly containing code-mixed words like *"YouTube views ko vote samjhne wale agar is bar Nahi jita to Kabhi Nahi jitega"*, the performance could be improved with better alignments and possibly a hybrid approach with minimal supervision.

6. Conclusion

This paper presents various approaches to sentiment analysis for Hinglish code-mixed tweets. Two different system architectures were researched: a supervised classification model incorporating cross-lingual embeddings for English-transliterated Hindi data and a transfer learning approach trained on English sentiment data and cross-lingual embeddings and applied to code-mixed data. Our results show that incorporating cross-lingual embeddings increases the performance from the baseline monolingual systems. In fact, the cross-lingual embeddings are so robust that even in a transfer learning setting, the system obtains an F1-score of *0.556*, which is comparable to the supervised classification scores of *0.606* and *0.616*.

As these were first experiments to apply cross-lingual embeddings for sentiment analysis for code-mixed Hinglish data, there is still a lot of room for improvement. First, we believe the cross-lingual embeddings can still be improved, as the embeddings constructed now are generic and can be further tailored with domain information to increase performance. In addition, the cross-lingual embeddings could also be post-processed with the monolingual embeddings to make them more robust and less susceptible to degradation. Additionally, more advanced classifiers like character-based convolution networks and Transformers, can be experimented with to produce better results out of the current embeddings. Finally, both the supervised and transfer learning approaches could be combined to further improve the results by providing multiple learning sources.

To conclude, we believe transfer learning incorporating cross-lingual embeddings is a viable approach to sentiment analysis for code-mixed data. As code-mixing is a common phenomenon in multilingual societies (Parshad et al., 2016), and the issue of transliteration exist in many South-Asian languages and other languages such as Arabic, the challenges addressed in this paper also hold for many other languages and tasks. As a result, the presented approach can be used for code-mixed text processing tasks in a variety of languages, and could be an important contribution to solve the data-acquisition bottleneck for NLP for code-mixed data.

7. Bibliographical References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Bakliwal, A., Arora, P., and Varma, V. (2012). Hindi subjective lexicon: A lexical resource for Hindi adjective polarity classification. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1189–1196, Istanbul, Turkey. European Language Resources Association (ELRA).

Bjerva, J. and Östling, R. (2017). Cross-lingual learning of semantic textual similarity with multilingual word representations. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 211–215, Gothenburg, Sweden, May. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chen, X. and Cardie, C. (2018). Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium, October-November. Association for Computational Linguistics.

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., and Weinberger, K. (2018). Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *CoRR*, abs/1710.04087.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Das, A. and Bandyopadhyay, S. (2010). SentiWordNet for Indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resources*, pages 56–63, Beijing, China.

Das, A., Chakraborty, T., Solorio, T., Gambäck, B., Aguilar, G., Kar, S., Garrette, D., and Pykl, S. (2020). Semeval-2020 task 9: Sentimix: Sentiment analysis for code-mixed social media text. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*. Association for Computational Linguistics.

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 462–471. Association for Computational Linguistics.

Gumperz, J. (1982). *Discourse Strategies*. Oxford University Press.

Hermann, K. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 58–68. Association for Computational Linguistics.

Joshi, A., Balamurali, A., and Bhattacharyya, P. (2010). A fall-back strategy for sentiment analysis in hindi: a case study. In *Proceedings of the 8th ICON*. Association for Computational Linguistics.

Joshi, A., Prabhu, A., Shrivastava, M., and Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 2482–2491, Osaka, Japan.

Labat, S. and Lefever, E. (2019). A classification-based approach to cognate detection combining orthographic and semantic similarity information. In G. Angelova, et al., editors, *Proceedings of Recent Advances in Natural Language Processing (RANLP 2019)*, pages 603–611, Varna, Bulgaria.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Luong, T., Pham, H., and Manning, C. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159. Association for Computational Linguistics.

Myers-Scotton, C. (1993). *Dueling Languages: Grammatical Structure in Code-Switching*. Claredon, Oxford.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Nguyen, D. and Seza Dogruoz, A. (2013). Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 857–862. Association for Computational Linguistics.

Parshad, R., Bhowmick, S., Chand, V., Kumari, N., and Sinha, N. (2016). What is India speaking? Exploring the "Hinglish" invasion. *Physica A: Statistical Mechanics and its Applications*, 449:375–389.

Pratapa, A., Choudhury, M., and Sitaram, S. (2018). Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 3067–3072. Association for Computational Linguistics.

Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K., and Maddila, C. (2017). Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Sharma, A., Gupta, S., Motlani, R., Bansal, P., Shrivastava, M., Mamidi, R., and Sharma, D. (2016). Shallow parsing pipeline - hindi-english code-mixed social media text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1340–1345. Association for Computational Linguistics.

Solorio, T. and Liu, Y. (2008a). Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 973–981. Association for Computational Linguistics.

Solorio, T. and Liu, Y. (2008b). Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 1051–1060. Association for Computational Linguistics.

Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany, August. Association for Computational Linguistics.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Choudhury, M. (2014). Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 974–979. Association for Computational Linguistics.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark, September. Association for Computational Linguistics.

Zhou, X., Wan, X., and Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412, Berlin, Germany, August. Association for Computational Linguistics.