CS-Embed at SemEval-2020 Task 9: The effectiveness of code-switched word embeddings for sentiment analysis

Frances A. Laureano De Leon
University of Birmingham
United Kingdom
laureanofa@gmail.com

Florimond Guéniat
Birmingham City University
United Kingdom
florimond.gueniat@bcu.ac.uk

Harish Tayyar Madabushi
University of Birmingham
United Kingdom
H.TayyarMadabushi.1@bham.ac.uk

Abstract

The growing popularity and applications of sentiment analysis of social media posts has naturally led to sentiment analysis of posts written in multiple languages, a practice known as code-switching. While recent research into code-switched posts has focused on the use of multilingual word embeddings, these embeddings were not trained on code-switched data. In this work, we present word-embeddings trained on code-switched tweets, specifically those that make use of Spanish and English, known as Spanglish. We explore the embedding space to discover how they capture the meanings of words in both languages. We test the effectiveness of these embeddings by participating in SemEval 2020 Task 9: *Sentiment Analysis on Code-Mixed Social Media Text*. We utilised them to train a sentiment classifier that achieves an F-1 score of 0.722. This is higher than the baseline for the competition of 0.656, with our team (codalab username *francesita*) ranking 14 out of 29 participating teams, beating the baseline.

1 Introduction

Sentiment Analysis (SA), or opinion mining, aims to identify subjective information from text. SA plays a pivotal role in promoting and identifying reception of products in platforms such as Amazon, or identifying the sentiment and opinions on numerous topics on social media platforms, such as Twitter or Facebook. As a consequence, SA is a popular task in the Natural Language Processing (NLP) community.

Sentiment analysis of social media posts has been a popular research topic, and has naturally led to the sentiment analysis of posts in languages other than English, as well as code-switched posts. Written text in which multiple languages co-exist, known as code-switching or code-mixing, is now more abundant thanks to the ample use of social media. Recent research on code-switched text has focused on using multilingual word embeddings, however, these embeddings were not trained using code-switched text. In this work, we present word-embeddings trained on code-switched social media text, which serves as our main contribution for SemEval 2020 Task 9 [1].

This task consisted of classifying tweets into one of three classes: positive, neutral, or negative. The dataset used for this task is the one provided by the competition organisers, who collected and annotated the corpus, (Patwa et al., 2020). We trained a BiLSTM (Schuster and Paliwal, 1997) classifier to detect the sentiment of code-mixed tweets by creating our own code-switched embeddings. Using the techniques outlined in section 3, large amounts of code-switched text can be collected to train code-switched embeddings for different NLP applications, such as semantic processing, dependency parsing (Schuster et al., 2019), as well as sentiment analysis, all of which have been identified as challenges posed by code-mixing in NLP tasks (Patwa et al., 2020). We release code for the code-switched sentiment classifier, crawled Twitter data, and associated experimental data, including hyper-parameters [2].

[1] https://competitions.codalab.org/competitions/20789

[2] https://github.com/francesita/CS-Embed-SemEval2020

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 922–927
Barcelona, Spain (Online), December 12, 2020.

2 Background

To our knowledge, there is a lack of literature and work on code-switched (CS) word embeddings. There is one paper on word embeddings using synthetic code-switched data (Pratapa et al., 2018). There is, however, considerably more literature on bilingual word embeddings created on parallel monolingual corpora. These are often used in cross-lingual and bilingual tasks (Zhou et al., 2015; Zhang et al., 2018).

Numerous methods have been introduced for creating bilingual word-embeddings. For instance, canonical correlation analysis (CCA) finds the association or correlation between two vectors, (Lu et al., 2015) to create bilingual word embeddings. Bilingual Canonical Correlation Analysis (BiCCA) was also developed, it is a technique in which monolingual embeddings are mapped to the same space by use of CCA, (M.Faruqui and C.Dyer, 2014). Not long after, BiCCA was extended to the learning of non-linear deep canonical correlation (DCCA), which often outperforms linear CCA (Lu et al., 2015).

This method was followed by the introduction of Bilingual Skip-gram, in which bilingual representations of words are learned from scratch, (Loung et al., 2015). Zhou et al. (2015) proposes learning bilingual sentiment word embeddings (BSWE) for English-Chinese SA by use of labelled documents and their translation rather than using parallel corpora with results that outperformed state of the art at the time.

Artexte et al. (2017) proposed to train embeddings individually, in their own monolingual space, and to map one monolingual embedding to the space of the other by use of a linear transformation. This technique makes use of bilingual dictionaries with as little as 25 words.

However, recent progress shows that bilingual methods of creating word embeddings are not well-suited for code-mixed tasks, (Pratapa et al., 2018). Indeed, there are grammatical structures that are not captured by monolingual text that exist when code-mixing occurs (J.M.Lipski, 2005; Sankoff and Poplack, 1981). It appears that bilingual embeddings created from monolingual or synthetic data may not be well suited for studies of code-mixed text. By training CS word embeddings, words that are usually used together in the same space when language switching occurs, will be clustered together (Jurafsky and Martin, 2019). We believe code-switched embeddings will be more robust in classifying CS text than others because the patterns in language when code-switching occurs will be captured in the embeddings. For these reasons, we have decided to train word embeddings on code-switched data.

3 System Overview and Experimental Set-up

We chose to train our own multilingual word embeddings using code-switched social media text because we hypothesise that these embeddings will be more effective in downstream tasks than combining monolingual embeddings from two different languages. To this end, we collected code-switched tweets in Spanglish [3]. The external tools utilised for this work are Gensim's word2vec model [4], natural language toolkit (NLTK) [5], keras version 2.2.4 [6], and tensorflow version 2.0 [7].

3.1 Twitter Data Collection

In order to train the code-switched word embedding[8], over 1,000,000 tweets were collected using an in-house code between 12 July, 2019 and 31 August, 2019 and between December 2019 and February 2020. Tweepy was used to gather the tweets with the Cursor function which allows for tweets to be searched for by query word, language, location and date.

First, a text document containing 315 words in Spanish and Spanglish was created and used as a code-switch key word list. The Spanish words chosen are part of a list of 500 most commonly used Spanish words according to the Dictionary of the Royal Spanish Academy. This list of 500 words was cross-referenced with a Portuguese dictionary in order to remove any words in common, as both these languages are romance languages. Words containing less than four letters were removed, to decrease the

[3] https://github.com/francesita/CS-Embed-SemEval2020

[4] https://radimrehurek.com/gensim/models/word2vec.html

[5] https://www.nltk.org/

[6] https://pypi.org/project/Keras/

[7] https://www.tensorflow.org/

[8] see https://github.com/francesita/Code-switch-Embeddings-for-Sentiment-Analysis

Table 1: SelEval2020 Task 9 Results

Rank	Users	Best Score
1	LiangZhao	0.806
2	rachel	0.776
3	asking28	0.756
14	**francesita (this work)**	**0.722**
	...	
23	suraj1ly (**organiser baseline**)	0.656

chances of overlap with other languages sharing words with Spanish, such as Tagalog. Proper nouns were removed as well. The Spanglish words included in the list refer to words that are hybrids of Spanish and English, such as the word *janguear* which both means and stems from the English phrase *hang-out* and ends with *-ar*, which allows the word *janguear* to be conjugated as a typical *-ar* infinitive Spanish verb. We looked-up the most commonly used Spanglish words in order to include these in the key word list. Noteworthy, many of the Spanglish key words included are typically used in Puerto Rico and Florida, which may induce bias, though efforts were made to include Spanglish words used by Hispanics living in Texas, California and other parts of the United States.

The words in the code-switched document were used as the query word in the cursor function in tweepy, while the language of the tweets extracted was English. Resulting collected tweets were considered code-mixed and were used to train the word-embeddings.

3.2 Preprocessing

NLTK was used for preprocessing the collected tweets. The tweets were tokenized. Spanish and English punctuation, as well as stop words were removed. We also extended contractions in English (such as can't to cannot). Emojis were kept, as these are useful for training sentiment classifiers (Lo et al., 2017).

3.3 Creation of Word-embeddings

Twitter data collected using methods described in 3.1 were used to train code-switched word-embeddings using Gensim's word2vec model, (Mikolov et al., 2013). We decided to train our embeddings using the CBoW algorithm rather than skip gram as it is better suited to smaller datasets, (Zhang et al., 2018). The embeddings were trained for 20 epochs with size 100, window of 5, and 10 workers for 30 iterations.

3.4 Description of BiLSTM Classifier

A BiLSTM model was used to classify the sentiment of the tweets. It was trained on 12,002 tweets, validated on 2,998 tweets and tested on 3,789 tweets. The BiLSTM model contained one embedding layer, in which the word embeddings we created were utilised. It had three bidirectional layers with dropout of 0.2, two dense layers, one with dimension 100 and dropout 0.3 and the other with dimension 3. The activation function used was *relu* for all layers except for the output classification layer, for which the *softmax* activation function was used. Adamax was used as the optimiser for the model, with learning rate 0.0002. Early Stopping was also used with min_delta 0.0002 and patience 5.

4 Results

As previously mentioned, our team decided to train word embeddings with code-switched text as we believed that these embeddings would be more effective than other multilingual embeddings for downstream tasks. Our team results and ranking in Task 9 are presented in table 1.

4.1 Embedding Evaluation

In the initial stages of data gathering to make code-switched word embeddings, the ten most similar words to each of the key words were used to do some simple evaluations on the embeddings. This

Table 2: Problematic Word Mappings

Word	Most Similar Words	Problem
calor	anitta(*I*), muitocalor(*port*), ozuna(*I*), muito(*port*), gosto(*port*)	language and current events
gracias	thanks(*en*), obrigada(*port*), muchas(*es*), grazie(*it*), bendiciones(*es*)	language
amazon	rainforest(*en*), fires(*en*), deforestation(*en*), wildfires(*en*), brazil(*en*)	current events

allowed us to identify some issues with early data collection methods, such as the inadvertent collection of code-switched tweets in Portuguese, Italian, and Tagalog mixed with English. Our embeddings were also biased to current events at the time of data collection. Table 2 shows examples of what issues were present in early embeddings. The letter in between brackets on the table indicate a language, or a pop culture icon (denoted by *I*).

We believe early issues were caused by the limit of tweet extractions set forth by Twitter (there is a 7 day limit the amount of data that can be collected at one time) as well as the limited number of key words used in early stages. The embeddings, after all, are only as good as the amount of data used to train them.

Happenings and current events of the time also influenced the embedding space to some degree as seen in table 3. Many of the words surrounding *gobernador* are related almost exclusively to Puerto Rico's government, which was having protests during the time of data collection. This shows that the time-frame in which data was collected allowed for bias to exist in the embeddings.

Table 3: Most Similar Words to *gobernador* (Governor)

Gobernador (governor)	renuncie (resign), gobernadora (female governor), renunciar (to resign), abajorosello (down with Rosello), cabron (expletive bastard/ cuckold)

Word embeddings may be biased in terms of what words are clustered together, which can be seen with the word *presidente* in table 4. It is evident that most of the words, and proper nouns clustered around the word *presidente* are associated to the political systems and leaders in the Americas.

Table 4: Examples of Bilingual Mapping from CS Embeddings

Word	Similar Words
Abuelo (grandfather)	abuela (grandmother), grandpa, abuelita (grandma), dad, grandma, cousin, grandparents
Novio (boyfriend)	esposo (husband), boyfriend, novia (girfriend), aver (to have), bf (boyfriend), gf (girlfriend)
Presidente (President)	president, dictator, presidentethe, presidentes, pres, guaido, reelection

We believe that as most of the code-switching between Spanish and English occurs in the Americas, the word embeddings reflect this. As we continue to gather data, much of the bias should disappear, as it is present due to the time-frame in which tweets were gathered as shown by table 5.

Table 5: Early and Later embeddings for the word *Calor*

Calor (Hot)	Most Similar Words
Early Embedding	anitta (artist), muitocalor (Portuguese), ozuna (artist), muito (Portuguese)
Latest Embedding	frío (cold), piscina (pool), sed (thirst), agua (water)

However, it is likely that not all bias will disappear, as code-switching occurs regionally, and data collected from these regions will likely show bias to that part of the world. For example, if code-switched data is collected from India, the concerns of the people, political or otherwise, are likely to be very different from the concerns of someone living in the Americas. Since code-switched word embeddings will be used for specific languages where code-switching is prominent, the embeddings need only reflect the happenings of that region.

4.2 Quantitative Analysis

We created a confusion matrix to understand where the system misclassified the sentiment of tweets. We utilised the development dataset as the labels for the test data are not yet released. As can be seen in table 6, neutral tweets were the most challenging to classify. These tweets were overwhelmingly classified as positive by our model. In the future, we plan on comparing the output of the classifier, and to use concurrent models to prevent miscalssification of neutral tweets into the positive class. It can also be seen that the negative class is well identified in our model, with no misclassified negative tweets, see table 6, and that the positive class is classified correctly for the most part. Results can be found Table 7.

Table 6: Confusion Matrix. Rows are actual classes while columns are predicted classes.

	Not Positive	Positive
True Not Positive	860	639
True Positive	457	1042
	Not Neutral	Neutral
True Not Neutral	1548	457
True Neutral	639	354
	Not Negative	Negative
True Not Negative	2492	0
True Negative	0	506

Our team also trained bilingual embeddings using a library of mutilingual unsupervised word embeddings (Lample et al., 2017) [9] to compare to our code-switched embeddings, with the only difference in the models and training procedure being the embeddings themselves. The results are presented in table 7. Results are comparable; it is noteworthy that our embeddings only contain 255,062 vocabulary words, compared to the 4,027,169 words in the bilingual embeddings. Both models had difficulties distinguishing the neutral class from the positive class, however, the model trained with code-switched embeddings was better at classifying positive tweets.

Table 7: Code-switched and Bilingual Embedding Comparison with dimension 100

	Precision		Recall		f-1 score		support
	CS	Bilingual	CS	Bilingual	CS	Bilingual	
Positive	0.62	0.62	**0.70**	0.68	**0.66**	0.65	1499
Neutral	**0.44**	0.43	0.36	0.36	0.39	0.39	993
Negative	1.00	1.00	1.00	1.00	1.00	1.00	506
macro avg	**0.69**	0.68	0.68	0.68	0.68	0.68	2998

5 Conclusion

In previous work, multilingual word embeddings have been used to classify code-switched text, however, these embeddings were not trained on code-switched data. In this work, we presented multilingual embeddings trained on code-switched social media text. [10] The results show that code-switched embeddings are able to capture the meanings of words when code-mixing occurs despite having a vocabulary much smaller in size. They also show that for the application of sentiment analysis, these embeddings are successfully employed to train a sentiment classifier, see table 1. In future work, we will use Google's Bidirectional Encoder Representations from Transformer (BERT) to obtain contextualised word embeddings from code-switched data. To our knowledge, although BERT has multilingual models, there are no code-switched models. We will also continue to collect data for code-switched word embeddings in Spanglish and begin collection for Hinglish.

[9] https://github.com/facebookresearch/MUSE

[10] https://github.com/francesita/Code-switch-Embeddings-for-Sentiment-Analysis

References

M. Artexte, G. Labaka, and E. Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume I: Long Papers)*, pages 451–462, July.

J.M.Lipski. 2005. Code-switching or borrowing? no sé so no puedo decir, you know. In *Selected proceedings of the second workshop on Spanish sociolinguistics*, pages 1–15.

D. Jurafsky and J.H. Martin. 2019. *Speech and Language Processing (3rd ed. draft).* 3rd edition edition.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043.*

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

M. Loung, H. Pham, and C. Manning. 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.

A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

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

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

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

D. Sankoff and S. Poplack. 1981. A formal grammar for code-switching. *Research on Language & Social Interaction*, pages 3–45.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *CoRR*, abs/1902.09492.

L. Zhang, S. Wang, and B. Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, January.

H. Zhou, L. Chen, F. SHi, and D. Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 430–440. Association for Computational Linguistics, July.