FII-UAIC at SemEval-2020 Task 9: Sentiment Analysis for CodeMixed Social Media Text using CNN

Lavinia Aparaschivei[1], Andrei Palihovici[1], Daniela Gîfu[1,2]
[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi
[2]Institute of Computer Science, Romanian Academy - Iasi Branch
{ lavinia.n.aparaschivei, andrei.r.palihovici, daniela.gifu}@info.uaic.ro

Abstract

The "Sentiment Analysis for Code-Mixed Social Media Text" task at the SemEval 2020 competition focuses on sentiment analysis in code-mixed social media text[1], specifically, on the combination of English with Spanish (Spanglish) and Hindi (Hinglish). In this paper, we present a system able to classify tweets, from Spanish and English languages, into positive, negative and neutral. Firstly, we built a classifier able to provide corresponding sentiment labels. Besides the sentiment labels, we provide the language labels at the word level. Secondly, we generate a word-level representation, using Convolutional Neural Network (CNN) architecture. Our solution indicates promising results for the Sentimix Spanglish-English task (0.744), the team, Lavinia_Ap, occupied the 9th place. However, for the Sentimix Hindi-English task (0.324) the results have to be improved.

1 Introduction

The explosive growth of social media (SM) platforms offers a new type of user, multilingual, that can alternate two languages or more in the same conversation (Udupa and Khapra, 2010). In linguistics, it is call Code-Switching (Adel *et al.*, 2013) or language alternation. Any virtual communication channel creates the possibility for people from different countries to share their sentiments without restrictions, about anything (Gîfu and Cioca, 2014), often using multiple languages. This explosion of sentiments has aroused the interest of many researchers for sentiment analysis (SA) for code-mixed SM message. Research on SA has focused, especially, on understanding the dynamics of sentiment in SM, most of them choosing Twitter because it provides free API very useful for data retrieval goal. That facility allows the developer to find real time tweets from different multilingual users. Actually, the analysis of tweets, which includes language alternation, is a challenging Natural Language Processing (NLP) issue. Code-mixing (CM) has several challenges to apply NLP techniques, such as word-level language identification or semantic processing (Myers-Scotton, 1993).

The goal of this paper is to implement a model for CM content on Twitter, which imply two objectives: first, the classification of positive, negative, and neutral tweets by generating word-level representation, using Convolutional Neural Network (CNN). The Spanish-English code-mixed content has become ubiquitous on the Internet, creating the need to process this form of natural language. A code-mixed sentence retains the underlying grammar and script of one of the languages it is comprised of.

The legitimate question of this survey is*: Can we achieve results comparable with those of our peers using more standard, less customized techniques?*

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 928–933
Barcelona, Spain (Online), December 12, 2020.

The rest of the paper is structured as follows: section 2 describes other works related to sentiment analysis for code-mixed social media text, section 3 presents the dataset and methods of this study, section 4 briefly relates the results we have obtained, followed by section 5 with the conclusions.

2 Background

Recent research regarding SA for code-mixing social media is becoming more attractive and challeng-ing (Patwa et al., 2020). The use of sentiment resources (Gîfu et al., 2014) has proven to be a neces-sary step for training and evaluating systems that implement SA, which also include fine-grained opin-ion mining (Balahur et al., 2010). A relevant work (Hu and Liu, 2004) is based on lexicon expansion techniques by adding synonymy and antonymy relations provided by WordNet (Miller and Fellbaum, 1998; Miller, 1993). In (Liu, Hu, and Cheng, 2005; Hu and Liu, 2004) an opinion lexicon was devel-oped, compounded by a list of positive and negative opinion words or sentiment words for English (around 6,800 words) and Spanish (around 1,500 words). A similar approach has been used for building WordNet (Strapparava and Valitutti, 2004) which expands six basic categories of emotions.

Several experiments have been performed on social media texts including code-mixed data. The first step toward information gathering from these texts is to identify the languages present. Until now, several language identification experiments or tasks have been performed on several codemixed language pairs such as Spanish-English (Goldbarg, 2009; Solorio *et al.*, 2011), French English (Voss *et al.*, 2014), Hindi-English (Bali *et al.*, 2014), Bengali-English (Mandal *et al.*, 2014). Many shared tasks have been organized for language identification of code-mixed texts, studied with promising results in other papers (Gopal *et al*, 2017, Heike *et al.,* 2013; Mishra *et al.*, 2018), as well. Language Identification in Code-Switched Data 5 was one of the shared tasks, which covered four language pairs such as Spanish-English, Modern Standard Arabic and Arabic dialects, Chinese-English, and Nepalese English. In the case of Indian languages, Mixed Script Information Retrieval (Royal *et al.*, 2015) shared task at FIRE-20156 was organized for eight code-mixed Indian languages such as Bangla, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, and Telugu mixed with English.

3 Dataset and Method

This section contains details about dataset built as part of SemEval-2020 Task 9 “Sentiment Analysis for Code-Mixed Social Media Text” and the study methodology, which was the basis for solving it.

3.1 Dataset

Regarding the combination of English with Spanish, the dataset consists from 18789 tweets, in CONLL format (1), and split in 3 parts: 12002 tweets for training, 2998 tweets for validation, and 3789 tweets for testing. Regarding the combination of English with Hindi, the dataset consists from 20000 tweets, in CONLL format (1), and split in 3 parts: 14000 tweets for training, 3000 tweets for validation, and 3000 tweets for testing, as in the example.

meta	uid	sentiment	
token	lang1_id		(1)
token	lang2_id		

Here, Uid label is a unique id for each tweet; lang1, lang2 labels correspond to the language pair language pair [here, Spanish (SPA) - English (ENG)], lang1 would be ENG and lang2 would be SPA.
We have three special labels, described below: (1) First is named ambiguous and it is used to tag words where the context surrounding that word is not clear enough to determine the language to which it belongs.

For example: the word a is a determiner in English and a preposition in Spanish.

(2) Second is called other and it is used to tag usernames, emoticons, symbols, punctuation marks, and other similar tokens that do not represent words. (3) Third is named ne and it is used to tag named entities, which are proper nouns.

In order to properly conduct an analysis of CS data, it must be identified correctly. It is a difficult work, considering the fact that the named entities are usually written the same, regardless of languages. To disambiguate these named entities, we need human annotators. It is a lot of work to do that includes defining absolute and correct guidelines for annotation (Molina *et al.*, 2019).

3.2 Method

This research presents a method able to classify tweets, written by Spanish users, specifically Spanish-English bilinguals, in three classes: positive, negative and neutral.

Baselines. We compare our approach with some common approaches as Support Vector Machines (SVM) and the TF-IDF representation with the combination of word unigrams and bigrams with F1-macro under 0,80.

Settings. We conduct the experiments building a classifier able to provide corresponding sentiment labels. Then, we generate a word-level representation, using Convolutional Neural Network (CNN) (see Figure 1). Besides the sentiment labels, we provide the language labels at the word level. The word-level language tags are EN (English), SPA (Spanish), HI (Hindi), mixed, and univ (e.g., symbols, @ mentions, hashtags).

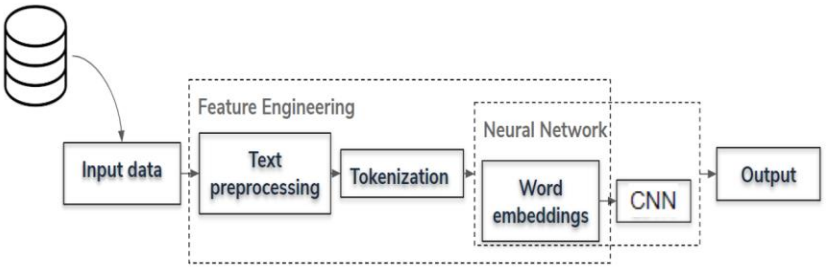

Figure 1: Overall Architecture

For Sentiment Analysis Spanish-English (SA_SPAN-ENG), the first objective was to classify tweets labelled with 'positive', 'negative', and 'neutral'. An example of a Spanish-English tweet can be seen below, with the lang label corresponding to the language to which the word belongs:

The	lang1
best	lang1
fall	lang1
is	lang1
..	other
Fall	lang1
in	lang1
LOVE	lang1
❤	other
Collar	lang2
rojo	lang2
$	other
14.90	other
Pedidos	lang2
096.880.7384	other
#neckless	lang1

#collar	lang2
#accesorios	lang2
http://t.co/6brIVHD2Xx	other

Once our data frame was created we pursued to the text preprocessing. In order to create a reliable dataset, we automatically striped the redundant information, like stop words and special characters using NLTK library. Given the fact, tweets contains informal text, word-level representation is a significant problem. A solution could be the character-level representation. In order to memorize aspects of word orthography the previous level takes characters as atomic units to derive the embedding (Joshi *et al.*, 2016). This is the reason for choosing CNN algorithm. This increases the robustness of the model, which is important for noisy social media data. Filters learn intermediate word feature representations during the convolution operation. The word-embedding layer learns jointly with the neural network model as the training takes place, while allowing us to gain a dense representation of the word vector spaces. The stack of Conv1D layers provides comparable results to other, more heavy-duty layers such as LSTM or recurrent ones at a fraction of the processing cost. The network was trained over 100 epochs on a personal laptop running Windows 10.

4 Results

Below, the official results for each individual subtask using the development and test sets are presented. We report Precision (P), Recall (R) and F1-score (F1), for each baseline on all classes.

Sub-Task 1_Spanish-English

Training Set			
Model	**P**	**R**	**F1**
CNN	0.823	0.981	0.895
Testing Set			
CNN	**0.820**	**0.979**	**0.892**

Table 1: Results for Spanish-English.

Sub-Task 2_Hindi-English:

Training Set			
Model	**P**	**R**	**F1**
CNN	0.315	0.351	0.332
Testing Set			
CNN	**0.310**	**0.340**	**0.324**

Table 2: Results for Hindi-English.

Our best result over the test dataset for SPAN-ENG achieved a P of 82% for the positive sentiments, an R of 0.979 and an F1-score of 0.892, but it is lower for HI-ENG. One of the reasons for the modest results for the HI-ENG dataset is that some EN words (e.g. 'costly') can be written in Hindi with different spelling variations. Note that, 30% of tokens were eliminated from the English-Spanish dataset, mostly filler words, such as stop words or misspelt words. The dataset is unbalanced - there are 37161 occurrences of English words throughout the entire dataset, with 57801 Spanish words, 150 ambiguous-labeled ones and tagged as 94 "mixed".

5 Conclusions

In this paper, we present description of the system that we have used in SemEval Task 9. With our model, we were able to achieve fifth position in Sentimix Spanglish-English task in evaluation. It combines CNNs networks, which prove to be very effective of training process of SA. We conduct several experiments on a real world code-mixed social media dataset and we have found that pre-processing steps played a huge role in increasing F1, especially for English-Spanish. For Hindi-English, it is necessary a large and diverse data collection. In order to improve the results, an approach regarding representing the word-vector space could be a good solution.

Reference

Balahur, A., Boldrini, E., Montoyo, A., and Mart´ınez-Barco, P. 2010. The OpAL sys-tem at NTCIR 8 MOAT. *Proceedings of NTCIR-8 Workshop Meeting*, 241–245.

Gîfu, D. and Cioca, M. 2014. *Detecting Emotions in Comments on Forums*. International Journal of Computers Communications and Control, I. Dzitac, F.G. Filip, M.-J. Manolescu (eds.), Vol. 9, no. 6/2014 (Dec.), Agora University Editing House, pp. 694-702.

Gîfu, D., Teodorescu, M., Ionescu, D. (2014). *Pragmatical Rules for Success*. International Letters of Social and Humanistic Sciences, vol. 26, pp. 18-28.

Gopal Patra, B., Das, D., and Das, A. 2018. *Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task@ ICON-2017*.

Heike A., Ngoc Thang V., Franziska K., Schlippe, T., and Schultz, T. 2013. Recurrent Neural Network Language Modeling for Code Switching Conversational Speech. *Proceedings of ICASSP* 2013.

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168-177.

Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., and Carman, M. 2016. Are word embedding-based features useful for sarcasm detection? *arXiv preprint arXiv:1610.00883*.

Kumar Lal, Y., Kumar, V., Dhar, M., Shrivastava, M., and Koehn, P. 2019. De-Mixing Sentiment from Code-Mixed Text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 371-377.

Liu, B., Hu, M., and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. *Proceedings of the 14th international conference on World Wide Web*, 342-351.

Mandal, S., Dipta Das, S., and Das, D. 2018. Language identification of bengali-english code-mixed data using character & phonetic based lstm models. *arXiv preprint arXiv:1803.03859*.

Mishra, P., Danda, P., and Dhakras, P. 2018. *Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches.*

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. 2019. *Overview for the second shared task on language identification in code-switched data*.

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

Myers-Scotton, C. 1993. *Common and uncommon ground: Social and structural factors in codeswitching. Language in society*, 22(4):475–503.

Negrón Goldbarg, R. 2009. Spanish-English Codeswitching in Email Communication. *Language@ internet*, 6(3).

Patwa, Parth, et al. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020),* Barcelona, Spain, December. Association for Computational Linguistics. 2020.

Prabhu, A., Aditya, J., Manish, S., and Vasudeva, V. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*.

Rudra, K., Rijhwani, S., Begum, R., Bali, K., Choudhury, M., and Ganguly, N. 2016. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1131-1141.

Sarkar, K. 2018. *JU_KS@ SAIL_CodeMixed-2017: Sentiment Analysis for Indian Code Mixed Social Media Texts*.

Sequiera, R., Choudhury, M., Gupta, P., Rosso, P., Kumar, S., Banerjee, S., and Chakma, K. 2015. Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval. *FIRE Workshops*, Vol. 1587, 19-25.

Solorio, T., Sherman, M., Liu, Y., Bedore, L. M., Peña, E., D., and Iglesias, A. 2011. Analyzing language samples of Spanish–English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17:3, 367-395.

Strapparava, C. and Valitutti, A. 2004. Wordnet Affect: an Affective Extension of Wordnet. *LREC*, 4:40, 1083-1086.

Udupa, R. and Khapra, M. M. 2010. Improving the Multilingual User Experience of Wikipedia Using Cross-Language Name Search. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 492–500.

Voss, C. R., Tratz, S., Laoudi, J., and Briesch, D. M. 2014. Finding Romanized Arabic Dialect in Code-Mixed Tweets. *LREC*, 2249-2253.