On the logistical difficulties and findings of Jopara Sentiment Analysis

Marvin M. Agüero-Torales
DECSAI, University of Granada
Granada, Spain
maguero@correo.ugr.es

David Vilares
Universidade da Coruña, CITIC
A Coruña, Spain
david.vilares@udc.es

Antonio G. López-Herrera
DECSAI, University of Granada
Granada, Spain
lopez-herrera@decsai.ugr.es

Abstract

This paper addresses the problem of sentiment analysis for Jopara, a code-switching language between Guarani and Spanish. We first collect a corpus of Guarani-dominant tweets and discuss on the difficulties of finding quality data for even relatively easy-to-annotate tasks, such as sentiment analysis. Then, we train a set of neural models, including pre-trained language models, and explore whether they perform better than traditional machine learning ones in this low-resource setup. Transformer architectures obtain the best results, despite not considering Guarani during pre-training, but traditional machine learning models perform close due to the low-resource nature of the problem.

1 Introduction

Indigenous languages have been often marginalized, an issue that is reflected when it comes to design natural language processing (NLP) applications, where they have been barely studied (Mager et al., 2018). One of the places where this is greatly noticed is Latin America, where the dominant languages (Spanish and Portuguese) coexist together with hundreds of indigenous languages such as Guarani, Quechua, Nahuatl or Aymara.

In this context, the Guarani language plays a particular role. It is an official language in Paraguay and Bolivia. Besides, it is spoken in other regions, e.g. Corrientes (Argentina) or Mato Grosso do Sul (Brazil), alongside with their official languages. Overall, it has about 8M speakers. Its coexistence with other languages, mostly Spanish, has contributed to its use in code-switching setups (Muysken, 1995; Gafaranga and Torras, 2002; Matras, 2020) and led to Jopara, a code-switching between Guarani and Spanish, with flavours of Portuguese and English (Estigarribia, 2015).

Despite its official status, there is still few NLP resources developed for Guarani and Jopara. Abdelali et al. (2006) developed a parallel Spanish-English-Guarani corpus for machine translation. Similarly, Chiruzzo et al. (2020) developed a Guarani-Spanish parallel corpus aligned at sentence-level. There are also a few online dictionaries and translators from Guarani to Spanish and other languages.[1] Beyond machine translation, Maldonado et al. (2016) released a corpus for Guarani speech recognition that was collected from the web; and Rudnick (2018) presented a system for cross-lingual word sense disambiguation from Spanish to Guarani and Quechua languages. There also are a few resources for PoS-tagging and morphological analysis of Guarani, such as the work by Hämäläinen (2019) and Apertium;[2] and also for parsing, more specifically for the Mbyá Guarani variety (Dooley, 2006; Thomas, 2019), under the Universal Dependencies framework.

In the context of sentiment analysis (SA; Pang et al., 2002; Liu, 2012), and more particularly classifying the polarity of a text as positive, negative or neutral, we are not aware of any previous work; with the exception of (Ríos et al., 2014). They presented a sentiment corpus for the Paraguayan Spanish dialect, which also includes words in English and Portuguese. However, there were few, albeit relevant, words of Guarani (70) and Jopara[3] (10), in comparison to the amount of the ones in Spanish (3,802) (Ríos et al., 2014, p. 40, Table II). Overall, SA has focused on rich-resource languages for which data is easy to find, even when it comes to code-switching setups (Vilares

[1] https://gn.wiktionary.org/, https://es.duolingo.com/dictionary/Guarani/, https://www.paraguay.gov.py/traductor-guarani, https://www.iguarani.com/, https://glosbe.com/gn, and Mainumby (Gasser, 2018).

[2] https://github.com/apertium/apertium-grn

[3] Tokens that mix n-grams of characters from Guarani and Spanish, e.g.: '*I understand*' would be '*entiendo*' (es), '*ahechakuaa*' (gn) and '*aentende*' (jopara).

Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pages 95–102

https://doi.org/10.26615/978-954-452-056-4_012

et al., 2016), maybe with a few exceptions such as English code-switched with languages found in India (Sitaram et al., 2015; Patra et al., 2018; Chakravarthi et al., 2020). In this context, although some previous work has developed multilingual lexicons and methods (Chen and Skiena, 2014; Vilares et al., 2017); for languages such as Guarani and other low-resource cases (where web text is scarce), it is hard to develop NLP corpora and systems.

Contribution Our contribution is twofold. First, we collect a corpus for polarity classification of Jopara tweets, which mixes Guarani and Spanish languages, being the former the dominating language in the corpus. We also discuss on the difficulties that we had to face when creating such resource, such as finding enough Twitter data that shows sentiment and contains a significant amount of Guarani terms. Second, we train a set of neural encoders and also traditional machine learning models, in order to have a better understand of how old versus new models perform in this low-resource setup, where the amount of data matters.

2 JOSA: The Jopara Sentiment Analysis dataset

In what follows, we describe our attempts to collect Jopara tweets. Note that ideally we are interested in tweets that are as Guarani as possible. However, Guarani is intertwined with Spanish, and thus we have focused on Jopara, aiming for Guarani-dominant tweets, in contrast to Ríos et al. (2014). We found interesting to report failed attempts to collect such data, since the proposed methods would most likely work to collect data in rich resource languages. We hope this can be helpful for other researchers interested in developing datasets for low-resource languages in web environments.

In this line, Twitter does not allow to automatically crawl Guarani tweets, since it is not included in its language identification tool. To overcome this, we considered two alternatives: (i) using a set of Guarani keywords (§2.1), and (ii) scrapping Twitter accounts that mostly tweet in Guarani (§2.2).

2.1 Downloading tweets using Guarani keywords - An unsuccessful attempt.

As the Twitter real-time streamer can deal with a limited number of keywords, we consider 50 different keywords which are renewed every 3 hours, and used them to sample tweets. To select such keywords, we considered two options:

1. *Dictionary-based keywords*: We used 5.1K Guarani terms from a Spanish-Guarani word-level translator.[4] We then downloaded 2.1M tweets and performed language identification with three tools: (i) `polyglot`,[5] (ii) `fastText`(Joulin et al., 2016) and (iii) `textcat`.[6] We assume that the text was Guarani if at least one of them classified the text as Guarani. After this, we got 5.3K tweets. Next, a human annotator was in charge of classifying such subset, obtaining that only 150 tweets, over the initial set of 2.1M samples, were prone to be Guarani-dominant.

2. *Corpus-based keywords*: We first merged two Guarani datasets[7] (Scannell, 2007), that were generated from web sources and included biblical passages, wiki entries, blog posts or tweets, among other sources. From there, we selected 550 terms, including word uni-grams and bi-grams with 100 occurrences or more. Again, we downloaded tweets using the keywords and collected 7M of tweets, but after repeating the language identification phase of step 1, we obtained a marginal amount of tweets that were Guarani-dominant.

Limitations This approach suffered from a low recall when it came to collect Guarani-dominant tweets, while similar approaches have worked when collecting data for rich-resource languages, where a few keywords were enough to succesfully download tweets in the target language (Zampieri et al., 2020). In this context, even if tweets contained a few Guarani terms, there were other issues: (i) words that have the same form in Spanish and Guarani such as '*mano*' ('*hand*' and '*to die*'), (ii) loanwords,[8] such as '*pororo*' ('*popcorn*') or '*chipa*' (traditional Paraguayan food, non-translatable); (iii) or simply tweets where the majority of the content was written in Spanish. Overall, this has been a problem experienced in other low-resource setups (Hong et al., 2011; Kreutz and Daelemans,

[4] https://github.com/SENATICS/traductor-espanhol-guarani

[5] https://polyglot.readthedocs.io/en/latest/Detection.html

[6] https://www.nltk.org/_modules/nltk/classify/textcat.html

[7] BCP-47 *gn* and *gug* codes.

[8] Frequent in Paraguay and border countries (Pinta, 2013).

2020), so we decided instead to look for alternatives to find Guarani-dominant tweets.

2.2 Downloading tweets from Guarani accounts - A successful attempt.

In this case, we crawled Twitter accounts that usually tweet in Guarani.[9] We scrapped them, and obtained more than 23K Guarani and Jopara tweets from a few popular users (see Appendix A.1). Using the same Guarani language identification approach as in 1, we obtained $8,716$ tweets. To eliminate very similar tweets that could contaminate the dataset, we removed tweets with a similarity greater than 60%, according to the Levenshtein distance. After applying this second cleaning step, we obtained a total of 3,948 tweets.

The dataset was then annotated by two native speakers of Guarani and Spanish. They were asked to: (i) determine whether the tweet was strictly written in Guarani, Jopara or other language (i.e., if the tweet did not have any words in Guarani); and determine whether the tweet was positive, neutral or negative. For sentiment annotations consolidation, we proceeded similarly to the SemEval-2017 Task 4 guidelines (Rosenthal et al., 2017, § 3.3).[10] We then filtered the corpus by language, including only those labeled as Guarani or Jopara, to ensure the samples are Guarani-dominant. This resulted into $3,491$ tweets.

Limitations Although this second approach is successful when it comes to collect a reasonable amount of Guarani-dominant tweets, it also suffers from a few limitations. For instance, the first part of Table 1 shows that due to the nature of the crawled Twitter accounts (who tweet about events, news, announcements, greetings, ephemeris, tweets to encourage the use of Guarani, etc.), there is a tendency to neutral tweets. Also, as the number of selected accounts was small, the number of discussed topics might be limited too. We comment on this a bit further in the Appendix A.1.

Balanced and unbalanced versions As we are interested in identifying sentiment in Jopara tweets, we also created a balanced version of JOSA. Note that unbalanced settings are also interesting and might reflect real-world setups. Thus, we will report results both on the unbalanced and balanced setups. More particularly, we split each corpus into training (50%), development (10%), and test (40%). We show the statistics in Table 1.

For completeness, in Table 2 we show for the balanced corpus the top five most frequent terms (we only consider content tokens) for Guarani, Spanish and some language-independent tokens, such emoticons. This was done based on a manual annotation of a Guarani-Spanish native speaker.

Version	Total	Positive	Neutral	Negative
Unbalanced	$3,491$	349	$2,728$	414
Balanced	$1,526$		763	

Version		Train	Development	Test
Unbalanced	$3,491$	$1,745$	349	$1,397$
Balanced	$1,526$	763	152	611

Table 1: JOSA statistics and splits for the unbalanced/balanced versions.

Category	#Terms	Most frequent
Guarani	$4,336$	guaranime, ñe'ẽ, mba'e, guarani, avei
Spanish	$1,738$	paraguay, guaraní, no, es, día
Other*	$1,440$	alcaraz, su, rt, juan, francisco
Mixing	368	guaraníme, departamento-pe, castellano-pe, castellanope, twitter-pe
Emojis	112	🇵🇾 😂 👏 xD :)

*We include reserved words, proper nouns, acronyms, etc.

Table 2: Frequent terms for the balanced JOSA.

3 Models

Due to the low-resource setup, we run neural models and pre-trained language models, but also other machine learning models, such as complement naïve Bayes (CNB) and Support Vector Machines (SVMs) (Hearst et al., 1998), since they are less data hungry, and could help shed some light about the real effectiveness of neural models on Jopara texts. In all cases, the selection of the hyperparameters was done over a small grid search based on the dev set. We report the details in the Appendix A.2.

Naïve Bayes and SVMs We tokenized the tweets[11] and represented them as a 1-hot vector of unigrams with a TF-IDF weighting scheme. We used Pedregosa et al. (2011) for training.

Neural networks for text classification We took into account neural networks that process in-

[9] We followed http://indigenoustweets.com/gn/. We did not use an external human annotator as in 1, since the crawled accounts tend to tweet in Guarani.

[10] We obtained a slight agreement following Cohen's kappa metric (Artstein and Poesio, 2008).

[11] We used the TweetTokenizer from the NLTK library.

put tweets as a sequence of token vector representations. More particularly, we consider both long short-term memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and convolutional neural networks (CNN) (LeCun et al., 1995), as implemented in NCRF++ (Yang and Zhang, 2018). Although the former are usually more common in many NLP tasks, the latter have also showed traditionally a good performance on sentiment analysis (Kalchbrenner et al., 2014).

For the input word embeddings, we tested: (i) randomly initialized word vectors, following an uniform distribution, (ii) and pre-trained non-contextualized representations and more particularly, FastText's word vectors (Bojanowski et al., 2017) and BPEmb's subword vectors (including the multilingual version, which supports Guarani) (Heinzerling and Strube, 2018). In both cases, we also concatenate a second word embedding, computed through a char-LSTM (or CNN).

Pre-trained language models We also fine-tuned recent contextualized language models on the JOSA training set. We tested BERT (Devlin et al., 2019) including: (i) beto-base-uncased (a Spanish BERT) (Cañete et al., 2020), and (ii) multilingual bert-base-uncased (mBERT-base-uncased, pre-trained on 102 languages). We also tried more recent variants of multilingual BERT, in particular XLM (Lample and Conneau, 2019). Note that BERT models use a wordpiece tokenizer (Wu et al., 2016) to generate a vocabulary of the most common subword pieces, rather than the full tokens, and that in the case of the multilingual models, none of the language models used considered Guarani during pre-training.

4 Experiments

Reproducibility The baselines and tweet IDs[12] are available at `https://github.com/mmaguero/josa-corpus`.

We run experiments for the unbalanced and balanced versions of JOSA, evaluating the macro-accuracy (to mitigate the impact of the neutral class in the unbalanced setup). Table 3 shows the comparison. Note that all models, even the non-deep-learning models, only use raw word inputs and do not consider any additional information or hand-crafted features,[13] yet they obtained results that are in line with those of more recent approaches.

	Corpus	
Model	**Unbalanced**	**Balanced**
CNB	0.50	0.55
SVM	0.55	0.54
^{C}CNN-^{W}BiLSTM	0.45	0.57
^{C}BiLSTM-^{W}CNN	0.49	0.53
$_{BPEmb,gn}$ ^{C}CNN-^{W}BiLSTM	0.46	0.53
$_{BPEmb,gn}$ ^{C}BiLSTM-^{W}CNN	0.42	0.50
$_{BPEmb,es}$ ^{C}CNN-^{W}BiLSTM	0.45	0.52
$_{BPEmb,es}$ ^{C}BiLSTM-^{W}CNN	0.45	0.50
$_{BPEmb,m}$ ^{C}CNN-^{W}BiLSTM	0.47	0.52
$_{BPEmb,m}$ ^{C}BiLSTM-^{W}CNN	0.43	0.48
$_{FastText,gn}$ ^{C}CNN-^{W}BiLSTM	0.46	0.53
$_{FastText,gn}$ ^{C}BiLSTM-^{W}CNN	0.42	0.51
$_{FastText,es}$ ^{C}CNN-^{W}BiLSTM	0.46	0.52
$_{FastText,es}$ ^{C}BiLSTM-^{W}CNN	0.46	0.46
BETO$_{base,uncased}$	**0.64**	**0.64**
mBERT$_{base,uncased}$	0.55	0.58
XLM-MLM-TLM-XNLI-15	0.46	0.49

^{C}Encodes character sequence. ^{W}Encodes word sequence.

Pre-trained embeddings are represented with a prefix together with their language ISO 639-1 code (except for m: multilingual).

Table 3: Experimental results on JOSA, both on the balanced and unbalanced setups.

With respect to the experiments with CNNs and BiLSTMs encoders, we tested different combinations using character representations, which output is first concatenated to a second external word vector (as explained in §3), and then fed to the encoder. Among those, the model that used a character-level CNN and a word-level BiLSTM encoder obtained the best results. Still, the difference with respect to traditional machine learning models is small. We hypothesize this might be due to the low-resource nature of the task. Finally, the pre-trained language models that use transformers architectures, in particular BETO, obtain overall the best results, despite not being pre-trained on Guarani. We believe this is partly due to the presence of Spanish words in the corpora and also to the cross-lingual abilities that BERT model might explode, independently of the amount of word overlap (Wang et al., 2019).

Error analysis on the balanced version of JOSA Figure 1 shows the confusion matrices for a representative model of each machine learning family (based on the accuracy): (i) CNB, (ii) the best BiLSTM-based model (CNN-BiLSTM), and (iii) Spanish BERT (BETO). There seems to be different tendencies in the miss-classifications that different models make. For instance, CNB tends to over-classify tweets as negative, while both deep

[12] Contact the authors for more details.

[13] In order to keep an homogeneous evaluation setup.

learning models show a more controlled behaviour when predicting this class. Although for the three models neutral tweets seem to be the easiest to identify, both deep learning models are clearly better at it. Finally, when it comes to identify positive tweets, BETO seems to show the overall best performance. These different tendencies indicate that an ensemble method could be beneficial for low-resource setups such as the ones that JOSA represent, since the models seem to be complementary to certain extent. In this context, we would like to explore this line of work in the future, following previous studies such as Jhanwar and Das (2018), which showed the benefits of combining different machine learning models for Hindi-English code-switching SA.

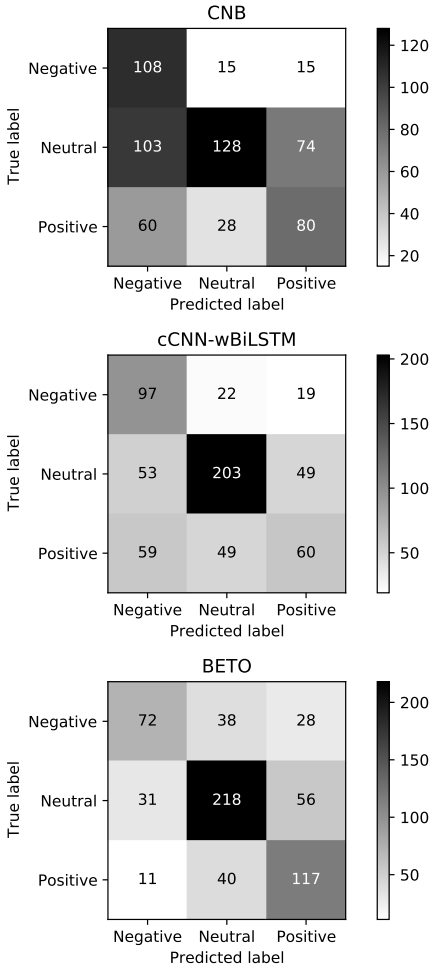

Figure 1: Confusion matrix for the balanced version of JOSA and the predictions of a representative member of each machine learning family: CNB, a BiLSTM-based model and Spanish BERT (BETO).

5 Conclusion

This paper explored sentiment analysis on Jopara, a code-switching language that mixes Guarani and Spanish. We collected the first Guarani-dominant dataset for sentiment analysis, and described some of the challenges that we had to face to create a collection where there is a significant number of Guarani terms. We then built several machine learning (naïve Bayes, SVMs) and deep learning models (BiLSTMs, CNNs and BERT-based models) to shed light about how they perform on this particular low-resource setup. Overall, transformers models obtain the best results, even if they did not consider Guarani during pre-training. This poses interesting questions for future work such as how cross-lingual BERT abilities (Wang et al., 2019) can be exploited for this kind of setups, but also how to improve language-specific techniques that can help process low-resource languages efficiently.

Acknowledgements

We thank the annotators that labelled JOSA. We also thank ExplosionAI for giving us access to the Prodigy annotation tool[14] with the Research License. DV is supported by a 2020 Leonardo Grant for Researchers and Cultural Creators from the FBBVA.[15] DV also receives funding from MINECO (ANSWER-ASAP, TIN2017-85160-C2-1-R), from Xunta de Galicia (ED431C 2020/11), from Centro de Investigación de Galicia 'CITIC', funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program) by grant ED431G 2019/01.

References

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski. 2006. Guarani: a case study in resource development for quick ramp-up mt. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,"Visions for the Future of Machine Translation*, pages 1–9.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[14] https://prodi.gy/

[15] The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert A Dooley. 2006. Léxico guarani, dialeto mbyá, com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística. e referências. *Cuiabá: Summer Institute of Linguistics*.

Bruno Estigarribia. 2015. Guarani-spanish jopara mixing in a paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222.

Joseph Gafaranga and Maria-Carme Torras. 2002. Interactional otherness: Towards a redefinition of codeswitching. *International Journal of Bilingualism*, 6(1):1–22.

Michael Gasser. 2018. Mainumby: un ayudante para la traducción castellano-guaraní. *arXiv preprint arXiv:1810.08603*.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lichan Hong, Gregorio Convertino, and Ed Chi. 2011. Language matters in twitter: A large scale study. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Madan Gopal Jhanwar and Arpita Das. 2018. An ensemble model for sentiment analysis of hindi-english code-mixed data. *CoRR*, abs/1806.04450.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665.

Tim Kreutz and Walter Daelemans. 2020. Streaming language-specific Twitter data with optimal keywords. In *Proceedings of the 12th Web as Corpus Workshop*, pages 57–64, Marseille, France. European Language Resources Association.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Diego Manuel Maldonado, Rodrigo Villalba Barrientos, and Diego P Pinto-Roa. 2016. Eñe'ẽ: Sistema de reconocimiento automático del habla en guaraní. In *Simposio Argentino de Inteligencia Artificial (ASAI 2016)-JAIIO 45 (Tres de Febrero, 2016)*.

Yaron Matras. 2020. *Language contact*. Cambridge University Press.

Pieter Muysken. 1995. Code-switching and grammatical theory. *The bilingualism reader*, pages 280–297.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: an overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Justin Pinta. 2013. Lexical strata in loanword phonology: Spanish loans in guaraní. Master's thesis, The University of North Carolina at Chapel Hill.

Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623.

Adolfo A Ríos, Pedro J Amarilla, and Gustavo A Giménez Lugo. 2014. Sentiment categorization on a creole language with lexicon-based and machine learning techniques. In *2014 Brazilian Conference on Intelligent Systems*, pages 37–43. IEEE.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Alexander James Rudnick. 2018. *Cross-Lingual Word Sense Disambiguation for Low-Resource Hybrid Machine Translation*. Ph.D. thesis, Indiana University.

Kevin P Scannell. 2007. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Dinkar Sitaram, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. 2015. Sentiment analysis of mixed language employing hindi-english code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 271–276. IEEE.

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France. Association for Computational Linguistics.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2016. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A Alonso. 2017. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Appendix

A.1 Twitter user accounts

We scraped the following Twitter user accounts and mentions: @ndishpy, @chereraugo, @Pontifex_grn, @lenguaguarani, @enga_paraguayo, @SPL_Paraguay, @rubencarlosoje1, as well as some keywords: '*guaranime*', '*avañe'ẽme*', '*remiandu*', '*#marandu*', '*reikuaavéta*', '*hesegua*', '*reheguápe*', '*rejuhúta*'.

Note that accounts such as @Pontifex_grn, @SPL_Paraguay and @lenguaguarani belong to influential people and organizations. For instance, the first belongs to Pope Franscico, the second to the Secretariat of Linguistic Policy of Paraguay, and the third is the account of the General Director of the 'Athenaeum of the Guarani Language and Culture'. On the other hand, the terms '*marandu*' (news) and '*remiandu*' (feeling, sense) are related to news, where the first term means 'news' or 'to report' and the second is the name of a Paraguayan newspaper section[16] that publishes in Guarani.

A.2 Hyperparameters search and implementation details

To set the machine learning baselines, two standard classifiers were chosen: a variant of Naïve Bayes, Complement Naïve Bayes (CNB) (Rennie et al., 2003) to correct the 'severe assumptions' made by the standard Multinomial NB classifier; and Support Vector Machine (SVM) using weighted classes, to mitigate the effect of unbalanced classes. For the CNB, we set $\alpha = 0.1$ and considered only unigrams, except for the balanced version, where the combined use of unigrams and bigrams showed more robust results. To train the SVMs, we tested different values for the kernels: the `sigmoid` kernel obtained the best results for the unbalanced version of JOSA, and the `poly` kernel obtained the best results for the balanced version.

We used the NCRF++ Neural Sequence Labeling Toolkit (Yang and Zhang, 2018) to train our deep learning models and the Hugging face package (Wolf et al., 2020) for the transformer-based models. Table 4 shows the hyper-parameters used to train these models, both for the unbalanced and balanced corpus. The pre-trained embeddings used for Spanish, Guarani (and also the multilingual ones) have 300 dimensions. Finally, we trained the CNN and BiLSTM models for 20 epochs with a batch size of 10, and the transformer-based models were trained for up to 40 epochs relying on early stopping (set to 3). To train the models we used a NVIDIA Tesla T4 GPU with 16GB.

Parameter	Options
Sklearn	
TF-IDF-Lowercase	[True, False]
TF-IDF-n-grams	[(1,1) - (3,3)]
SVM-Kernel	[poly, sigmoid, linear, rbf]
CNB-alpha	[1.0, 0.1]
NCRF++	
Optimizer	[Adam, AdaGrad, SGD]
Avg. batch loss	[True, False]
Learning rate	[5e-5 - 0.2]
Char hidden dim.	[100, 200, 400, 800]
Word hidden dim.	[50, 100, 200]
Momentum	[0.0, 0.9, 0.95, 0.99]
LSTM Layers	[1, 2]
Hugging Face	
Eval. steps	[200]
Eval. strategy	[steps]
Disable tqdm	[False]
Eval. batch size	[16, 32]
Train batch size	[16, 32]
Learning rate	[2e-5 - 3e-5*]
Dropout	[0.1 - 0.6]
Epoch	[30 - 40]
Weight decay	[0.0 - 0.3]

*Except for the multilingual models, where 5e-5 was necessary to converge.

Table 4: Hyperparameters for the training of the models, both for the unbalanced and balanced corpus.

[16] https://www.abc.com.py/especiales/remiandu/