Idiap and UAM Participation at MEX-A3T Evaluation Campaign

Esaú Villatoro-Tello[a,b], Gabriela Ramírez-de-la-Rosa[b], Sajit Kumar[c], Shantipriya Parida[b] and Petr Motlicek[b]

[a]*Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico*
[b]*Idiap Research Institute, Rue Marconi 19, 1920, Martigny, Switzerland*
[c]*Centre of Excellence in AI, Indian Institute of Technology Kharagpur, West Bengal, India*

Abstract

This paper describes our participation in the shared evaluation campaign of MexA3T 2020. Our main goal was to evaluate a Supervised Autoencoder (SAE) learning algorithm in text classification tasks. For our experiments, we used three different sets of features as inputs, namely classic word n-grams, char n-grams, and Spanish BERT encodings. Our results indicate that SAE is adequate for longer and more formal written texts. Accordingly, our approach obtained the best performance (F = 85.66%) in the fake-news classification task.

Keywords

Supervised Autoencoders, Text Representation, Deep Learning, Natural Language Processing

1. Introduction

In this era where social media and instant messaging is widely used for communication, the reach and volume of these text messages are enormous. The use of aggressive language or dissemination of false news is widespread across these communication channels. It is impossible to verify the text messages manually. We need automated systems that help users of these communication channels to determine if they are reading real or fake news or to try to flag when someone has been targeted with aggressive messages.

Besides the fact that most of the previous works done in these two tasks, namely aggressiveness detection and fake-news detection, are for English, little research has been done for Spanish using the most recent NLP techniques such as deep learning approaches. On the one hand, for aggressiveness detection, in past editions of the MEX-A3T[1] challenge [1], only three out of nine approaches used some deep learning classifier, particularly for CNN, LSTM, and GRU, with no good performances [2]. On the other hand, most of the current research on fake-news detection has been done for the English language, using graph CNNs [3], and more recently attention mechanism-based transformer models [4].

Our participation at MEX-A3T 2020 aimed at exploring the use of Supervised Autoencoder (SAE) [5] in two different text classification tasks: *i)* aggressiveness detection in Spanish tweets, where documents are very short and informal texts; and, *ii)* fake-news detection from Spanish newspapers,

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: evillatoro@correo.cua.uam.mx, esau.villatoro@idiap.ch (E. Villatoro-Tello); gramirez@correo.cua.uam.mx (G. Ramírez-de-la-Rosa); kumar.sajit.sk@gmail.com (S. Kumar); shantipriya.parida@idiap.ch (S. Parida); petr.motlicek@idiap.ch (P. Motlicek)
ORCID: 0000-0002-1322-0358 (E. Villatoro-Tello); 0000-0003-3387-6300 (S. Parida)

[1] https://sites.google.com/view/mex-a3t/home

Table 1
Features as inputs for the Supervised Autoencoder Method.

Features type	Sub-type	Identifier
Word n-grams	n=(1,2) and n=(1,3)	W
Char n-grams	n=(1,2) and n=(1,3)	C
BETO	*min*, *max*, and *mean* pooling	B
Word n-grams and Char n-grams		W+C
BETO and Word n-grams		B+W
BETO and Char n-grams		B+C
BETO, Word n-grams and Char n-grams		B+W+C

where documents are larger and contain a more formal written style. We found that SAE can generalize well for both tasks, particularly, for the aggression detection our approach obtains an F1 macro of 80.7%, while for the fake-news detection we reached the best score with an F1 macro of 85.6%.

2. Methodology

For both tasks, we aimed at evaluating the impact of recent generalization techniques, namely SAE [5] with a varied set of features as input vectors. Although SAE has been extensively evaluated in image classification tasks [6], very few works exist evaluating the impact of SAE in text classification tasks, e.g. language detection [7]. Next, we briefly describe the SAE theory, and we provide some details on how the document representation was generated for all the explored features.

2.1. Supervised Autoencoder

An autoencoder (AE) is a neural network that learns a representation (encoding) of input data and then learns to reconstruct the original input from the learned representation. The autoencoder is mainly used for dimensionality reduction or feature extraction [5]. Normally, it is used in an unsupervised learning fashion, meaning that we leverage the neural network for the task of representation learning. By learning to reconstruct the input, the AE extracts underlying abstract attributes that facilitate accurate prediction of the input.

Thus, an SAE is an autoencoder with the addition of a supervised loss on the representation layer. The addition of supervised loss to the autoencoder loss function acts as a regularizer and results in the learning of the better representation for the desired task [6]. For the case of a single hidden layer, a supervised loss is added to the output layer and for a deep supervised autoencoder, the innermost (smallest) layer would have a supervised loss added to the bottleneck layer that is usually transferred to the supervised layer after training the autoencoder.

For all our performed experiments, the overall configuration of the SAE model was done using nonlinear activation function (ReLU) with 3 hidden layers, the number of nodes in the representation layer was set to 300, and we trained to a maximum of 100 epochs.

2.2. Input Features

The SAE receives as input the representation of the document build using Spanish pre-trained BERT encodings (BETO [8]), traditional text representation techniques such as word and char n-grams (ranges 1-2 and 1-3), and, combinations of BETO encodings plus traditional words/char n-grams vectors.

We choose to evaluate the impact of word and char n-grams since as previous research has shown [9, 10, 11], word n-grams are capable of capturing the identity of a word and its contextual usage, while character n-grams are additionally capable of providing an excellent trade-off between sparseness and word's identity, while at the same time they combine different types of information: punctuation, morphological makeup of a word, lexicon and even context. For generating this type of features we used the `CountVectorizer` and `TfidfTransformer` libraries from the scikitlearn[2] toolkit. For the case of the fake-news detection task, we empirically chose the best values for the `min-df` and `max-df` parameters, which are reported on Table 3. For the aggressiveness task, these values were fixed (for all the experiments) to `min-df=` 0.001 and `max-df=` 0.3.

Additionally, we evaluate the impact of transformer-based models [12] as a language representation strategy. For our experiments we tested BETO[3], a BERT model trained on a large dataset of Spanish documents [8]. As known, the [CLS] token acts an "aggregate representation" of the input tokens, and can be considered as a sentence representation for many classification tasks [13]. Accordingly, we apply the following approaches for generating the representation of the document: *i)* for the aggressiveness task, each tweet is directly passed to the BETO model, and is represented using the encoding of the last hidden layer from the [CLS] token; ii) for the fake-news detection task, we split the news document into smaller chunks, obtain the [CLS] encoding of each chunk, and then we apply either a *min*, *max*, *mean* pooling for generating the final document representation. Table 1 depicts the type and variations of features tested during the training phase.

Finally, it is worth mentioning that we did not apply any preprocessing steps in any of the tasks. To validate our experiments, we performed a stratified 10 cross-fold validation strategy.

3. Aggressiveness Identification

The offensive language in Mexican Spanish corpus used for this task has 10,475 Spanish tweets. The training partition contains 7332 tweets with two possible classes (aggressive or non-aggressive). More details of this corpus can be found in [14]. Table 2 shows the results obtained in both, the validation phase and our two runs submitted for the final evaluation of this task over 3143 unseen tweets. The difference between the two submitted outputs, i.e., run id 1 and 2 (†), is the classifier, submission 2 was trained using a Multi-Layer Perceptron (MLP).

4. Fake-News Identification

The fake-news Spanish corpus used in this task has 971 news from 9 different topics. The training partition provided for the development stage has 676 news with a binary class (fake or true). Each news is compose by the headline, body, and the URL from where the news was published (the complete description of this corpus can be found in [15]). For our experiments, we used only the headline and the body of the news as a single document. Table 3 shows the results obtained in the development stage of the challenge, and the two runs submitted for the final evaluation of the tasks over 295 unseen news.

[2] https://scikit-learn.org/stable/index.html
[3] https://github.com/dccuchile/beto

Table 2
Results in validation and test phases reported in F-score for aggressive (F+), non-aggressive (F-), and macro average of the F-score (Fm).

	Validation phase			*Test phase*			
Input features	Fm	F+	F-	ID	Fm	F+	F-
W (1,2)	0.783	0.698	0.868	-	-	-	-
W (1,3)	0.777	0.690	0.864	-	-	-	-
C (1,2)	0.726	0.601	0.850	-	-	-	-
C (1, 3)	0.778	0.689	0.866	-	-	-	-
B (LHL)	0.742	0.628	0.856	-	-	-	-
C (1, 3) + W (1,2)	0.780	0.702	0.857	-	-	-	-
B + W (1,2)	0.787	0.694	0.879	-	-	-	-
B + C (1,3)	0.780	0.684	0.876	-	-	-	-
B + W (1,2) + C (1,3)	**0.803**	0.716	0.889	1	0.807	0.725	0.888
B + W (1,2) + C (1,3)†	**0.798**	0.702	0.894	2	0.801	0.706	0.895
Bi-GRU (baseline-given by track organizers)					0.798	0.712	0.884
BOW-SVM (baseline-given by track organizers)					0.777	0.676	0.878
Best system (in the task [1])					0.859	0.799	0.919

Table 3
Results in validation and test phases reported in F-score for fake-news (F+), real-news (F-), and macro average of F-score (Fm).

		Validation phase			*Test phase*			
Input features	`min-df,max-df`	Fm	F+	F-	ID	Fm	F+	F-
W(1,2)	0.01, 0.5	0.775	0.793	0.758	-	-	-	-
W(1,3)	0.01, 0.5	0.778	0.798	0.758	-	-	-	-
C(1,2)	0.01, 0.5	0.697	0.719	0.674	-	-	-	-
C(1, 3)	0.01, 0.5	0.757	0.768	0.745	-	-	-	-
B(min-pooling)		**0.843**	0.842	0.845	2	**0.856**	**0.844**	**0.868**
B(max-pooling)		0.830	0.830	0.830	-	-	-	-
B(mean-pooling)		0.833	0.831	0.835	-	-	-	-
C(1, 3)+W(1,2)	0.01, 0.5	0.805	0.807	0.802	-	-	-	-
B+W(1,2)	0.01, 0.3	**0.845**	0.846	0.844	1	0.850	0.840	0.859
B+C(1,3)	0.01, 0.3	0.834	0.834	0.835	-	-	-	-
B+W(1,2)+C(1,3)	0.01, 0.3	0.833	0.831	0.835	-	-	-	-
B+W(1,2)+C(1,3)	0.01, 0.5	**0.848**	0.846	0.850	-	-	-	-
Third best system (in the track)						0.817	0.819	0.817
BOW-RF (baseline-given by track organizers)						0.786	0.785	0.787

5. Conclusions

This paper describes Idiap & UAM participation at the MEX-A3T 2020 shared task on the Classification of Fake-News and Aggressiveness analysis. Our participation aimed at analyzing the performance of recent generalization techniques, namely deep supervised autoencoders. To this end, we performed a comparative analysis among simple transformers based language representation strategies and traditional text representations such as word and character n-grams. Notably, the SAE method benefits the most when it is feed with input features generated from the combination of BERT encodings and word/char n-grams. Particularly, for the aggression detection task, our proposed approach can obtain

a relative improvement of 1.1% over the stronger baseline, while for the fake-news detection task the improvement over the baseline is 8.1%.

As future work, we plan to perform an analysis of what are the dataset characteristics that allow the SAE approach to provide good performances. Also, we want to evaluate the impact of SAE's hyperparameter tuning through optimization methods, such as Bayes Optimizer[16], and evaluate our proposed approach on other similar classification tasks.

Acknowledgments

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: "SM2: Extracting Semantic Meaning from Spoken Material" funding application no. 29814.1 IP-ICT and EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE), grant agreement: 833635. The first author, Esaú Villatoro-Tello is supported partially by Idiap, SNI-CONACyT, CONACyT project grant CB-2015-01-258588, and UAM-C Mexico during the elaboration of this work.

References

[1] M. E. Aragón, H. Jarquín, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, J.-P. Posadas-Durán, Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish, in: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain, September, 2020.

[2] M. E. Aragón, M. Á. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, D. Moctezuma, Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, 2019.

[3] F. Monti, F. Frasca, D. Eynard, D. Mannion, M. M. Bronstein, Fake news detection on social media using geometric deep learning, arXiv preprint arXiv:1902.06673 (2019).

[4] M. Qazi, M. U. S. Khan, M. Ali, Detection of fake news using transformer model, in: 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2020, pp. 1–6.

[5] Q. Zhu, R. Zhang, A classification supervised auto-encoder based on predefined evenly-distributed class centroids, arXiv preprint arXiv:1902.00220 (2019).

[6] L. Le, A. Patterson, M. White, Supervised autoencoders: Improving generalization performance with unsupervised regularizers, in: Advances in Neural Information Processing Systems, 2018, pp. 107–117.

[7] S. Parida, E. Villatoro-Tello, S. Kumar, P. Motlicek, Q. Zhan, Idiap submission to swiss-german language detection shared task, in: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS), 2020.

[8] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: to appear in PML4DC at ICLR 2020, 2020.

[9] Z. Wei, D. Miao, J.-H. Chauchat, R. Zhao, W. Li, N-grams based feature selection and text representation for chinese text classification, International Journal of Computational Intelligence Systems 2 (2009) 365–374.

[10] A. Kulmizev, B. Blankers, J. Bjerva, M. Nissim, G. van Noord, B. Plank, M. Wieling, The power of character n-grams in native language identification, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 382–389.

[11] F. Sánchez-Vega, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, E. Stamatatos, L. Villaseñor-Pineda, Paraphrase plagiarism identification with character-level features, Pattern Analysis and Applications 22 (2019) 669–681.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[14] M. J. Díaz-Torres, P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes-y Gómez, J. Aguilera, L. Meneses-Lerín, Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 132–136. URL: https://www.aclweb.org/anthology/2020.trac-1.21.

[15] J.-P. Posadas-Durán, H. Gomez Adorno, G. Sidorov, J. Moreno, Detection of fake news in a new corpus for the spanish language, Journal of Intelligent Fuzzy Systems 36 (2019) 4869–4876. doi:10.3233/JIFS-179034.

[16] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, 2012, pp. 2951–2959.