# User-Emotion Analysis leveraging Pre-trained Word Embedding Models for Albanian

An Empirical Evaluation

Marjana Prifti Skënduli
Department of Computer Science
University of New York Tirana
Albania
marjanaprifti@unyt.edu.al

Doriela Grabocka
Department of Computer Science
University of New York Tirana
Albania
dorielagrabocka@unyt.edu.al

## ABSTRACT

In our research, we leverage the analysis performed on micro-blogging texts in Albanian language, which enables the use of technologies to monitor and follow the feelings and perception of the people with respect to products, issues, events, etc. Our approach to emotion analysis tackles the problem of classifying a text fragment according to a set of pre-defined emotions, aiming to detect the emotional state of the writer conveyed through the text. We present a rigorous analysis of the emotion classification task performance with regards to different word representations from static (shallow) word embeddings to dynamic (deep contextualized) word embeddings and different machine learning classifiers from classical (shallow) Machine Learning models to Deep Learning models. Experimental evaluations show that best results have been obtained from Logistic Regression and Random Forest with regards to statistical supervised (shallow) models, while the CNN model with pre-trained fastText word embeddings reported the best overall results out of all selected models across the six benchmarking datasets, with an accuracy ranging from 90.67% - 97.49%. We also report experimental results of our extrinsic evaluator (emotion analysis task) on five word embedding models, pointing out that XLM-RoBERTa outperforms the other pre-trained multilingual LMs, among deep contextualized word embeddings. While fastText significantly outperforms among static word embedding models, obtaining an average of 9.86% accuracy improvement over the word2vec model.

## KEYWORDS

NLP, Emotion Analysis, word embedding, language model

## 1 Introduction

Recently, Deep Learning models have achieved state-of-the-art results on many NLP tasks, yet these models are trained from scratch, requiring large datasets, and abundant time to converge. On a second note, word embeddings have also played a significant role in numerous research problems in NLP. They serve as a warm start to the system and also steer models to richer and more insightful learning as opposed to the naïve one-hot encodings. Previously used frequency-based embedding techniques, like TF-IDF and co-occurrence vector, have been superseded by prediction-based embedding techniques, like Word2Vec, Glove and fastText. Emergence of contextualized embedding techniques, like BERT, is another major break-through for downstream NLP tasks. This motivates our work towards leveraging and assessing the impact of the whole range of word embeddings and pre-trained language models (LMs) in fine-grained emotion analysis. In light of the indisputable benefits of transfer learning, we decided to compare several pre-trained word embeddings and LMs, including but not limited to word2vec, fastText, BERT [1], XLM [2] and XLM-RoBERTa [3]. We hypothesize that their significant improvement can be gained for Albanian language too, in spite of being a low-resource language. Our work is the first comparative study conducted on pre-trained LMs on downstream NLP tasks in Albanian language.

## 2 Approach

While fully recognizing the scarcity of Albanian linguistic resources, we decided to build our own corpus and extract from there several benchmarking datasets. We fetched around 60K Facebook posts belonging to 119 shortlisted Albanian politicians, that were further manually annotated. The linguistically noisy and domain specific nature of the Facebook content, required us to elaborate a pragmatic yet comprehensive preprocessing workflow in order to ease the training comprehension of our models.

## 3 Experimental setup

Experiments were performed along two main perspectives:
(i) Exploiting different word representations from static word embeddings to dynamic (contextualized) word embeddings and (ii) evaluating different machine learning classifiers from classical Machine Learning models to Deep Neural Network models.

## 4 Conclusions

Deep Learning architectures have hitherto provided state-of-the-art performance due to their peculiarities to learn with small intervention on the data representation and feature engineering. Indeed, language models that leverage transfer learning are conceptually simple and empirically powerful, thus result-wise they achieve state-of-the-art accuracy.

In this work, we show for the first time the strong impact of cross-lingual language model pre-training on NLP tasks such as text classification. This empirical study demonstrates also, that rich and supervised/unsupervised pre-training is crucial for resource-constrained languages like Albanian.

## REFERENCES

[1] Devlin et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (May 2019). Retrieved from https://arxiv.org/abs/1810.04805.
[2] Lample, G. and Conneau, A. 2020. Cross-lingual Language Model Pretraining. arXiv.org. https://arxiv.org/abs/1901.07291.
[3] 2020. Arxiv.org. https://arxiv.org/pdf/1911.02116.pdf .