Multi-Language Neural Network Model with Advance Preprocessor for Gender Classification over Social Media Notebook for PAN at CLEF 2018

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma and Vitor Beires Nogueira
Computer Science Department, University of Évora, Portugal
{kshyp, tcg, pq, vbn}@uevora.pt

Basic Ideas

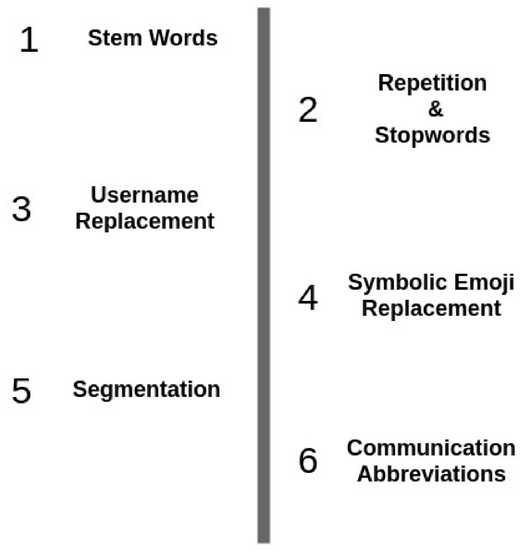
The idea is to treat tweet text as a collection of the dictionary words and then using indexing with simple dense architecture. But before making entry to the dictionary, preprocessing of tweets is done followed by text representation and construction of the classification model.

Principal Objective

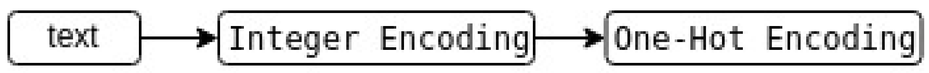
1. Data Preprocessing.
2. Text Representation.
3. System Modeling.
4. Discussion.

1 Data Preprocessing

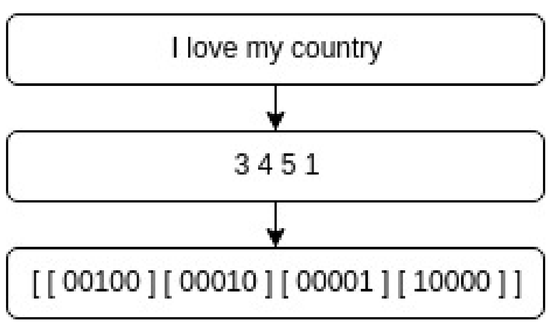
Here, the different preprocessing steps are discussed which are attained to build the input features for the machine learning algorithm.



2 Text Representation



Assuming the English dictionary ”country”: 1, ”very”: 2, ”I”: 3, ”love”: 4, ”my”: 5, considering the following figure:

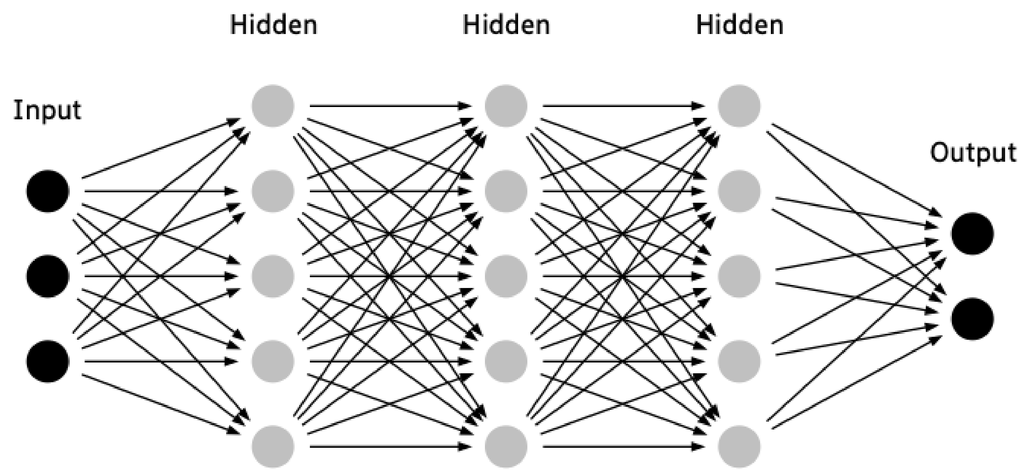


3 System Modeling

4 Dense System Architecture

After several iterations, the number of layers for the architecture was set to three.

- 1^{st} - layer - 1024 nodes - Relu
- 2^{nd} - layer - 512 nodes - Sigmoid
- 3^{rd} - layer - 256 nodes - Softmax



5 Main Result

Table 1 presents the results provided by the PAN 2018 organizing committee for the systems described in the previous section.

Language	Accuracy
English	72.79
Spanish	64.36
Arabic	72.20

Tabela 1: Results for PAN 2018 Author Profiling Task

6 Conclusions

- Simple Dense Architecture, Beyond Average Performance
- Data Preprocessing, Dictionary Creation

7 Limitation

- Only Suitable to Small Dataset, Dictionary Index out of Bound
- Binary Conversion Consumes too much of RAM

8 Further Work

- For unseen words, neighboring or similar word could be used
- Including Semantic Meaning
- Part of Speech (POS)
- Entity Extraction (EE)

Acknowledgements

The authors would like to thank COMPETE 2020, PORTUGAL 2020 Programs, the European Union, and LISBOA 2020 for supporting this research as part of Agatha Project SI & IDT number 18022 (Intelligent analysis system of open of sources information for surveillance/crime control) made in collaboration with the University of Évora. The colleagues Madhu Agrawal, Silvia Bottura Scardina and Roy Bayot provided insight and expertise that greatly assisted the research.