ADVANCED MACHINE LEARNING TECHNIQUES TO IDENTIFY EMOTIONS IN TEXTS

Atif Ali
Research Management Centre (RMC), Multimedia University, Cyberjaye 63100 Malaysia.
dralexaly@gmail.com

Zulqarnain Farid
Dept of Criminology, University of Karachi, Pakistan
discovercrimegene@gmail.com

Abstract: This research aims to learn to classify short texts (opinions) generated on the social network Twitter according to their feelings, applying advanced machine learning techniques such as neural networks. In the first stage, text classification was explored using an LSTM neural network. In the current stage, some ways of representing texts are being analyzed to create a corpus of embedded words that will be used in other experiments.

***Keywords*:** emotions, machine learning, neural networks, Twitter.

1. INTRODUCTION

A large amount of information in social media has led the scientific community to dedicate great efforts to analyzing, structuring, and processing this information. These media often express diverse opinions and feelings about society, products, services, politics, celebrities, etc. Companies, organizations, and governments are interested in knowing users' opinions about their activities. Digital marketing is mainly based on the opinions expressed on social networks.

Polarity detection in textual opinion is a widely researched task, especially for English. Identifying the emotion expressed in a the opinion is a less-researched task with ample research possibilities. In this regard, the frequently used approaches are supervised learning, which uses large amounts of text as input to the algorithms, and a dictionary of words associated with one or more emotions. The research group addressed these types of learning in the previous project.

II. Related Work

Deep learning approaches have demonstrated their ability to solve tasks related to natural language processing and artificial intelligence applications. Neural networks are effective when performing tasks of classifying texts [3]. [4], they apply a hybrid method with convolutional neural networks and recurrent neural networks to polarity classification in tweets. [5] uses a combination of artificial neural networks to recognize emotions in texts. The word embeddings technique consists of representing words as vectors of real numbers on which it is possible to carry out operations and obtain surprising results. [6] This is used to increase the effectiveness of classifying emotions.

Before the learning stage, it is necessary to carry out preprocessing actions to eliminate those characteristics that can produce noise in the following stages, for example:

- Tokenization and lemmatization
- Elimination of stopwords
- Elimination of images, links and references to users

Generally, social network users often use emojis to highlight what it wants to express, as a form of voice intonation or body language. In the previous project, it was shown that keeping emojis and hashtags is relevant. Therefore, they must be transformed into text. These types of elements, in general, are not considered in the works cited above.

Python is one of the most accepted programming languages by the scientific community. It is powerful and is characterized by its simplicity, open-source distribution and the possibility of integrating with multiple libraries. For text processing with Python, there are a variety of computer tools. In [7], some of them were analyzed, showing that Freeling and Stanford are the most reliable in terms of tokenization and grammatical labelling.

III. RESEARCH AND DEVELOPMENT LINE

This research project proposes to detect feelings expressed in texts, particularly textual opinions issued in a social network. The project is developed in the following stages:

- Review of the literature relevant to the problem of opinion and sentiment mining.
- Evaluation and comparison of deep learning techniques for text classification.
- Evaluation and comparison of other machine learning techniques for textual opinion classification.
- Development of a prototype for the classification of opinions.

The tweets captured for the previous project and others obtained last year totalled more than 150,000. Many of these tweets were discarded for containing only images, icons or text with little information for analysis.

IV. RESULTS OBTAINED/EXPECTED

In the first part of the project, experiments with neural networks for learning text classification began.

To combat the vanishing gradient, which occurs in Recurrent Neural Networks (RNN), LSTM (Long Short-Term Memory) networks arise, which are a special type of recurrent network [8]. The main characteristic of LSTMs is that the information can persist by introducing loops in the network diagram to

decide the next one [9]. Figure 1 shows the typical structure of this type of neural network.

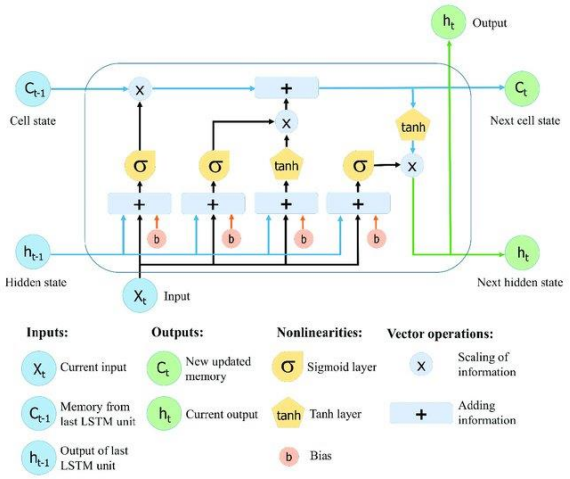

Figure 1. Long Short-Term Memory [9]

The main difference with traditional neural networks (RNT) is that they do not have the persistence (memory) of the previous data. An RNT cannot use its reasoning about previous events to decide on later ones. The dataset from the previous project [2], a set of tweets in Spanish classified by expressed emotions, was used to perform the LSTM network tests. for a first approach to the process, it was decided to work with two categories, in such a way to classify the tweets by their polarity. They were grouped as follows:

Positive = happiness and surprise. Negative = disgust, anger, sadness and fear

The valence of tweets in the EI-reg and EI-of datasets is displayed in Figure 2. Observe that, as expected, the chosen query terms resulted in the anger, fear, and sadness datasets having a majority of negative tweets and the pleasure dataset having a majority of positive tweets. The reason why the two affect dimensions are not perfectly connected is revealed by such tweets (or perfectly inversely correlated).

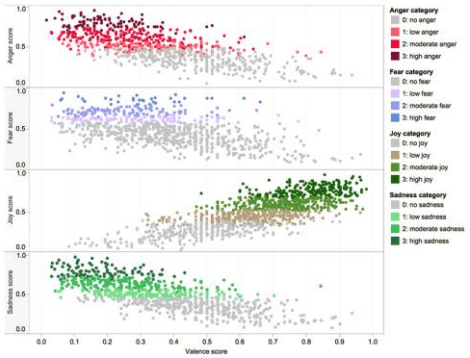

Figure 2: Valence of tweets in the EI-reg and EI-oc datasets. Shows the distribution of tweets according to the emotions expressed

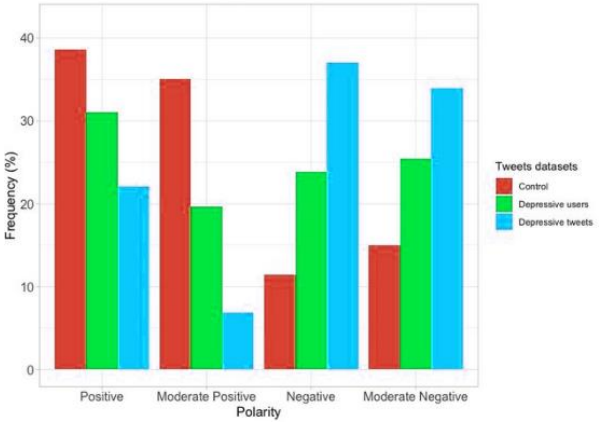

Figure 3. Polarities of the tweets according to the SentiCo Polarity tool in the three datasets.

The configuration and parameterization of the RNN were carried out using the Colaboratory tool provided by Google. Figure 3 contains part of the Python code used to configure the RNN

```
# RNN/LSTM is defined
defRNN():
      inputs =
Input(name='inputs',shape=[max_len])
      layer =
Embedding(max_words,120,input_length=max_l in)(inputs)
      layer = LSTM(64)(layer)
      layer = Dense(1,name='out_layer')
(layer)
      layer = Activation('softmax')(layer)
      model =
Model(inputs=inputs,outputs=layer)
      returnmodel
```

Figure 3. RN configuration

The process continues with the execution of the defined RNN and the evaluation of the built model, taking the training set separated in a previous stage as reference. Experimentation with neural networks is currently continuing.

The next objectives to be achieved are:

- Create a corpus of word embeddings from the collection of available tweets
- Experience learning with the corpus of word embeddings generated and with others available for research.
- Select the algorithms that best classify opinions.

This line of research is expected to continue and broaden the understanding of natural language processing. It is intended that this project encourage interest in research and this subject in the students.

V. Conclusion

With cutting-edge machine learning techniques like neural networks, this project intends to learn how to categorise brief words (opinions) created on the social network Twitter according to their emotions. The initial step investigated text classification using an LSTM neural network. In the current stage, several representational strategies are being examined to build a

corpus of embedded words that will be utilised in additional experiments.

References

[1]. Ali, A., Said, R. A., Rizwan, H. M. A., Shehzad, K., & Naz, I. (2022, February). Application of Computational Intelligence and Machine Learning to Conventional Operational Research Methods. In *2022* International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-6). IEEE.

[2]. Ali, A., Qasim, M., Dilawar, M. U., Khan, Z. F., Jadoon, Y. K., & Faiz, T. (2022, February). Nanorobotics: next level of military technology. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-7). IEEE.

[3]. A deep reinforcement learning approach to solve the vehicle routing problem with resource Constra... (2023). https://doi.org/10.2514/6.2023-2662.vid

[4]. Deep reinforcement learning. (2022). *The Science of Deep Learning*, 229-250. https://doi.org/10.1017/9781108891530.017

[5]. A. Ali et al., "The Threat of Deep Fake Technology to Trusted Identity Management," 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ICCR56254.2022.9995978.

[6]. Vanneschi, L., & Silva, S. (2023). Artificial neural networks. *Natural Computing Series*, 161-204. https://doi.org/10.1007/978-3-031-17922-8_7.

[7]. Introduction to semantics of programming languages. (2021). *Concepts and Semantics of Programming Languages 1*, 15-33. https://doi.org/10.1002/9781119824121.ch2.

[8]. Okafor, N., Delaney, D., & Mathew, U. (2022). ProxySense: A novel approach for gas concentration estimation using long short-term memory recurrent neural network (LSTM-RNN). https://doi.org/10.36227/techrxiv.20306418.v1.

[9]. Larsen, K. R., & Becker, D. S. (2021). Why use automated machine learning? *Automated Machine Learning for Business*, 1-22. https://doi.org/10.1093/oso/9780190941659.003.0001.