Sentiment Analysis through Conversational Data

Alexander Espronceda Gómez, Satu Elisa Schaeffer

Facultad de Ingeniería Mecánica y Eléctrica

alexander.esproncedago@uanl.edu.mx

Abstract

In this project, open-sourced software is proposed, which interprets the text entered by a person and determines how they are feeling at the moment, with the purpose of being used in tandem with another software or algorithms focused on conversational data. The study method used will make a comprehensive analysis of neural networks, as well as pattern recognition and data collection. The algorithm is open-source so anyone can add or remove modules as needed.

1. Introduction

Human beings are social beings, this is widely known. To survive, we must band together and communicate with each other, bonding in the process. This is thanks to a neural process called *empathy*, which is defined as a process that happens in our brains.

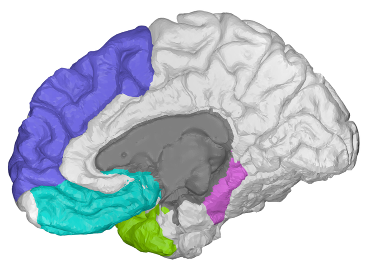

Figure 1: Lateral brain map of the parts in charge of the empathy processes. Drawing generated using BrainPainter [1].

Theoretically, a machine could be taught to process signals of distress and react accordingly using a machine learning algorithm.

2. Justification

This project could prove especially useful towards being used in projects designed for people who have trouble discerning when to console someone or having an idea of how other people or even themselves feel, such as the case of people with Asperger's Syndrome or other forms of high-functioning autism. To this end, the decision was made to work on this project.

3. Hypothesis

The hypothesis of this thesis is that using supervised machine learning with a neural network could accurately classify the sentiment behind an input text as "Good", "Neutral" or "Bad", with the purpose of being implemented in tandem with another software or algorithms focused on conversational data.

4. Objectives

The objective of this project is to make software capable of determining how the person that writes the input text is feeling according to the words in it, while keeping the code open-source so it can be used in other projects. This could be achieved thanks to the technology present in machine learning algorithms and an extensive amount of datasets.

5. Background

5.1 Basic Concepts

Machine Learning Also known as ML. The type of algorithm needed for automatic processing, making the machine "learn" (hence the name) over time given enough data.

Neural Network A Machine Learning algorithm that uses weights and filters to output data.

Natural Language Processing This is the method used for the algorithm to understand the content of the sentences, this is usually achieved by using tokenization but a preset corpus can also be used.

Sentiment Analysis This involves a ML algorithm, usually a Neural Network, that is able to analyze sentences and classify them according to the words used.

Corpus Preset internal dictionary that the algorithm uses.

Tokenizing Process that converts every word in the lexicon to an assigned number (called *token*) for easier processing.

5.2 Supervised Machine Learning

Supervised ML can be described, broadly and figuratively speaking, as a black box where some data is inserted as an input and numbers come out of it as an output [2].

This output, as opposed to other types of Machine Learning, is later analyzed and compared to real life data.

5.3 Sentiment Analysis

- The sentence to analyze is broken down to its component parts, this process is called *tokenization*, and the resulting products are called *tokens*.
- Every token is then tagged, making it part of an internal dictionary or *lexicon*
- A score is assigned to every token depending on the used dataset.

5.4 Tokenization

Taking the following example:

This is an example text

We can tell there are 5 words in the example phrase. So the tokenizing process would make the example look in the following way:

1, 2, 3, 4, 5

The interesting part about this process would happen if we used another example phrase, like the following:

This is another example

If we did the tokenization process, it would be processed in this way:

1, 2, 6, 4

6. Comparison to Related Work

There are several papers on similar projects, the following table marks the specifics of each one's relevant features compared to this project's.

Table 1: Comparison between existing literature and the present work: ✓ indicates the fulfillment of a criterion, otherwise × is used.

Project	Neural Network	Text Processing	Sentiment Analysis	Open Source	Modular
Blenn et al. [3] Maximum Entropy	✓	✓	✓	×	×
Blenn et al. [3] Support Vector Machines	✓	✓	✓	×	×
Blenn et al. [3] Lingpipe	✓	✓	✓	×	×
Morris et al. [4]	✓	✓	✓	×	×
Bird et al. [5]	✓	✓	×	✓	×
Pang et al. [6]	✓	✓	✓	✓	×
Ahmad et al. [7]	✓	✓	×	✓	×
Wang et al. [8]	✓	✓	✓	✓	×
Capuano et al. [9]	✓	✓	✓	×	×
Chiril et al. [10]	✓	✓	×	✓	×
Röchert et al. [11]	✓	✓	✓	✓	×
The present work	✓	✓	✓	✓	✓

7. Proposed Solution

This project is built on Python v3.8.10, The libraries used for this project to come to fruition are TensorFlow v2.6.0 and Keras v2.6.0 for the Neural Network section and Natural Language Toolkit v3.5 (also known as NLTK) for the tokenization and stemming process.

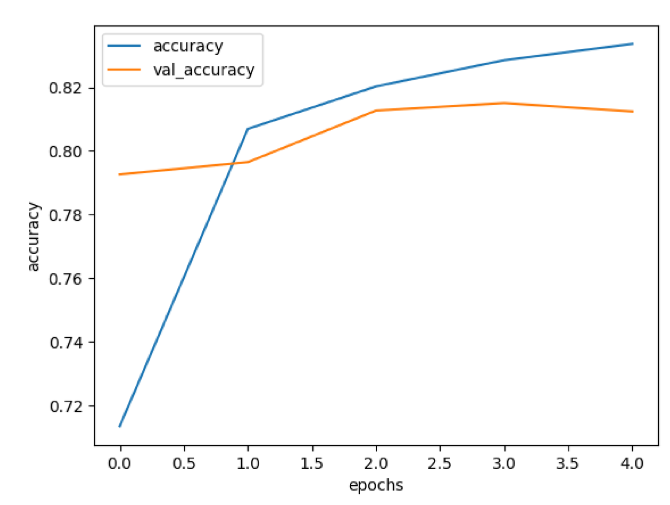

Figure 2: Accuracy values of the finished project

8. Experiments

The purpose of these experiments is to determine if the parameters chosen for this project are optimal and, if not, correct them and know the reason behind the improvement.

Lower loss and higher accuracy are preferred.

- Experiment 1: Base Experiment
- Experiment 2: More Datasets With Reduced Data Scope
- Experiment 3: Augmented LSTM Units
- Experiment 4: Augmented Datasets Without Reduced Data Scope
- Experiment 5: Reduced Classification Scope
- Experiment 6: Reduced Epochs
- Experiment 7: Added Stop Words
- Experiment 8: Extra Stop Words and Reduced Classification Scope

Table 2: Experimental results

	Training		Cross-Validation	
	Loss	Accuracy	Loss	Accuracy
Experiment 1	0.6916	0.7130	**0.8709**	0.6234
Experiment 2	0.5956	0.7576	0.7649	0.6821
Experiment 3	0.5829	0.7564	0.7373	0.6780
Experiment 4	0.5455	0.7741	0.6704	0.7110
Experiment 5	0.6222	0.6550	0.7186	0.5357
Experiment 6	0.6041	0.7451	0.6555	0.7097
Experiment 7	0.6030	0.7421	0.6579	0.7156
Experiment 8	0.3624	0.8337	0.3871	**0.8124**

9. Conclusions

These results determine that the proposed hypothesis is partially true: given enough data, a Machine Learning algorithm can learn to classify feelings and react accordingly, effectively learning how to identify patterns to an extent. However, high quality and volume data is needed for this to be reliable. Something that was only partially obtained for this project.

Overall, we can determine that, with the use of more consistent data, a favorable result can be achieved with the model used in this project.

In the future, this project would greatly benefit from a dataset that takes into consideration sentences that can be said in any context and still be correctly classified. And, of course, the less ortographical errors there are, the better.

References

[1] Razvan Marinescu, Arman Eshaghi, Daniel Alexander, and Polina Golland. Brainpainter: A software for the visualisation of brain structures, biomarkers and associated pathological processes. *arXiv preprint arXiv:1905.08627*, 2019.

[2] Xian-Da Zhang. Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence*, pages 223–440. Springer, 2020.

[3] Norbert Blenn, Kassandra Charalampidou, and Christian Doerr. Context-sensitive sentiment classification of short colloquial text. In Robert Bestak, Lukas Kencl, Li Erran Li, Joerg Widmer, and Hao Yin, editors, *NETWORKING 2012*, pages 97–108, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-30045-5.

[4] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *J Med Internet Res*, 20, Jun 2018. ISSN 1438-8871. doi:10.2196/10148.

[5] Jordan J Bird, Anikó Ekárt, and Diego R Faria. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–16, 2021.

[6] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

[7] Shakeel Ahmad, Muhammad Zubair Asghar, Fahad M Alotaibi, and Irfanullah Awan. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, 9:1–23, 2019.

[8] Bingkun Wang, Ning Wang, and Zhongsheng Chen. Research on air quality forecast based on web text sentiment analysis. *Ecological Informatics*, 64, 2021. ISSN 1574-9541. doi:https://doi.org/10.1016/j.ecoinf.2021.101354.

[9] Nicola Capuano, Luca Greco, Pierluigi Ritrovato, and Mario Vento. Sentiment analysis for customer relationship management: An incremental learning approach. *Applied Intelligence*, 51:3339–3352, 2021.

[10] Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31, 2021.

[11] Daniel Röchert, German Neubaum, and Stefan Stieglitz. Identifying political sentiments on youtube: A systematic comparison regarding the accuracy of recurrent neural network and machine learning models. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pages 107–121. Springer, 2020.

Online Poster

Online Repository