**6ᵗʰ PPgSI's Dissertations Workshop
2019**

# Depression Detection Model based on Brazilian Portuguese Facebook posts

| | |
|---|---|
| Author: | Felipe Morine Magami |
| Advisor: | Luciano Antonio Digiampietri |
| Co-advisor: | |

| Research lines: | [ ] Systems Development and Management | [ x ] Systems Intelligence |
|---|---|---|

| Research areas: | [ ] Database | [ ] Software engineering | [ x ] Artificial intelligence | [ ] Graphics processing |
|---|---|---|---|---|
| | [ ] Information technology management | [ ] Human-Computer Interaction | [ x ] Pattern recognition | [ ] Optimization |

| Application areas: | [ ] Enterprise environments / Business processes | [ ] Bioinformatics | [ ] Biometrics | [ ] Mobile devices |
|---|---|---|---|---|
| | [ ] Economy | [ ] Education / Distance learning | [ ] E-government | [ x ] Internet / Social Networks |
| | [ ] Games / Serious games | [ x ] Linguistics / Natural Language | [ ] Cheminformatics | [ ] Robotics |
| | [ ] Health | [ ] Other   Which? _____ | | [ ] General* |

| Period in the program (at the workshop date): | [ ] 2ⁿᵈ semester | [ x ] 3ʳᵈ semester | [ ] 4ᵗʰ semester | [ ] 5ᵗʰ semester |
|---|---|---|---|---|

| Qualifying: | [ ] Qualifying held in: dd/mm/yyyy | [ x ] Plan for qualifying in: 01/11/2019 |
|---|---|---|
| Defense: | Deadline for deposit: 01/02/2021 | Plan for defending in: 02/12/2020 |

| Publications associated with the master's project: | **No publications to date.** |
|---|---|

## The research project summary

**Context:**

According to the World Health Organization, in 2017, more than 300 million people suffered from depression, the leading cause for disability worldwide. Brazil, in particular, has one of the biggest rates of depression, with more than 11.5 million depressives. Identifying depressive social media users through their online posts and publications could lead to an improvement of current diagnosis services, helping depressive people seeking adequate treatment.

**Research problem:**

Though several works regarding social media and depression detection were found in our literature review, we could not identify one that tackled the problem for social media publications written in Brazilian Portuguese.

**Research objective:**

Our objective is to develop and evaluate models for depression detection based on the textual contents from Facebook posts written in Brazilian Portuguese. The best performing model should show better results than the proposed baseline.

**Characteristics of the proposed solution:**

The dataset to be utilized in this work consists of Facebook posts from 688 Brazilians. The volunteers answered a self-evaluating questionnaire widely used for depression diagnosis, and then were separated into four depression levels. We intend to employ several methods and techniques used for depression detection from text data found in the literature, in particular, methods for feature extraction, adapted to Portuguese, and the classification models that achieved the best results.

**Theoretical foundations:**

For feature extraction, we consider four approaches: lexicon usage, specifically the Portuguese version of the LIWC; Portuguese Part-of-Speech taggers; bag-of-words and n-grams representations; and Portuguese word embedding models. For classification purposes, we consider neural networks, support vector machines and Bayesian algorithms.

**Correlated works:**

- CHOUDHURY, M. et al. Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media. Cambridge, EUA: AAAI, 2013.
- TADESSE, M. M. et al. Detection of depression-related posts in Reddit social media forum. IEEE Access, IEEE, v. 7, p. 44883–44893, 2019.
- TROTZEK, M.; KOITKA, S.; FRIEDRICH, C. M. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering, IEEE, p. 1–1, 2019.

**Validation**

The accuracy, precision, recall and f-score from the models will be measured. Statistical tests will also be employed for comparison purposes. Because no benchmark or baseline can be used (since no work regarding Brazilian Portuguese has been found), the majority class from the dataset will be adopted as the baseline.

**Limitations, risks, and threats:**

Since the focus of this research is text-based data for the Portuguese language from social media posts, and given the lack of a public dataset with these characteristics, this work will only utilize the dataset previously described. Furthermore, this work does not refer to the related but distinct tasks of regression (as in, predicting a depression score instead of a level) and early prediction.

**Scientific contribution:**

We expect that our model, if verified its positive performance, will serve as a first step to a more automatic depression diagnosis service, especially for Brazilians. In addition, we believe that the understanding of text processing for Portuguese social media data will be increased.

**Technical contribution (if pertinent):**

A model for detection of depression based on text-based data from social media.

| The research method | |
|---|---|
| Genre (choose ONE) | [ ] Theoretical research    [ x ] Practical research    [ ] Empirical research    [ ] Methodological research |
| Nature (choose ONE) | [ ] Basic research        [ x ] Applied research |
| Approach (choose ONE) | [ x ] Quantitative research    [ ] Qualitative research    [ ] Quali-quanti research |
| Literature review* (you can choose more than one) | [ x ] Narrative review    [ ] Meta-analysis    [ ] Theoretical review <br> [ ] Descriptive review    [ x ] Qualitative systematic review    [ ] Realistic review <br> [ ] Scoping review    [ ] Umbrella review    [ ] Critical review |
| Main technical procedure (choose ONE) | [ x ] Experimental research    [ ] *Survey*    [ ] Ethnographic research <br> [ ] Bibliographic research    [ ] Case study    [ ] Grounded theory <br> [ ] Documental research    [ ] Participatory research    [ ] Design science <br> [ ] *Ex-post-facto* research    [ ] Research-action    [ ] Other Which?_____ |
| Data analysis (you can choose more than one) | [ x ] Descriptive statistics    [ ] Statistical test    [ ] Discourse analysis <br> [ ] Inferential statistics    [ ] Content analysis    [ ] Others: _____ |

* Definition of types of literature reviews established by Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

**Next steps**:
1. Complete the literature review;
2. Extract features from the dataset, according to the mentioned strategies;
3. Apply the selected classifiers;
4. Execute the created models and obtain the results;
5. Compare the results according to statistical tests;
6. Analyze the results;
7. Write the dissertation.

**Optional:** Provide a graphic layout that shows aspects of your research. For example: a flow-chart for building your solution or an infographic for your research proposal. If necessary, use the fourth page.