ANALYSIS OF SOCIAL MEDIA POSTS RELATED TO POSTPARTUM DEPRESSION A SUMMARY PROTOCOL ON HOW TO DEVELOP A REMOTE LABORATORY

Ulfeta Marovac*,
Aldina Avdić

Department for Technical Sciences,
State University of Novi Pazar,
Novi Pazar, Serbia

Abstract:

Analysing the content of posts from social networks related to specific health problems can contribute to improving the health of the general population. This study gives an analysis of posts related to postpartum depression, which was performed to automatically detect content that correlates with postpartum depression. Machine learning methods can be used to detect posts that correlate with postpartum depression. The specificity of the language in which the posts are written reduces the availability of training corpora and processing tools. In this paper, a topic analysis is provided and a model for the prediction of postpartum depression in posts using a corpus composed of posts from the Reddit and ana.rs forums is presented.

Keywords:

Social media, Postpartum depression, Machine learning, Topic analysis.

INTRODUCTION

Caring for women in the postpartum period is very important, but due to the care of the baby, women often neglect their physical and mental health. Very often, health problems are tried to be solved through advice on social networks.

How do recognize the state of postpartum depression? This question should be answered by psychiatrists and psychologists. Postpartum depression signs and symptoms may include: depressed mood or severe mood swings, excessive crying, difficulty bonding with the baby, withdrawal from family and friends, loss of appetite or eating much more than usual, inability to sleep (insomnia) or sleeping too much, overwhelming fatigue or loss of energy, reduced interest and pleasure in usual activities, intense irritability and anger, fear of unsuccessful parenting, hopelessness, feelings of worthlessness, shame, guilt or inadequacy, etc.

Correspondence:
Ulfeta Marovac

e-mail:
umarovac@np.ac.rs

DOI: 10.15308/Sinteza-2022-14-19

To determine postpartum depression (PPD), the Edinburgh scale is used, with a score greater than 12 indicating postpartum depression. EPDS is one of the most commonly used scales for assessing depressive symptoms of women who have given birth. The respondent estimates the weight for ten different depressive symptoms in the past seven days on a scale of 0 to 3 [1]. The possible range of results is from 0 to 30, where higher scores refer to more difficult symptomatic.

Detection of posts in which the symptoms of postpartum depression are described is one of the goals of this paper. Such research has already been done, but for posts in English. However, people express themselves most sincerely in their mother tongue, so it is necessary to adapt these analyses to other languages as well. In this paper, we will perform an analysis of posts in English and posts in Serbian translated into English. The presented methodology uses the tools available for the English language and adapts their use to the Serbian language.

The second section presents an overview of similar research related to the application of machine learning in predicting postpartum depression in fasts. The third section provides an overview of the materials used and the method. An analysis of the topic of the posts as well as the results of the methods for predicting postpartum depression in the posts are given in the fourth section. Finally, a conclusion is given and ideas for further research are presented.

2. RELATED WORK

In the era of the Internet and the social network of Facebook, Twitter, and Reddit, conditions have been created to collect large amounts of textual data through which it is possible to monitor personality behavior in posts on social networks. Thus, the context of the analysis of depression-related chatter on Twitter to glean insight into social networking about mental health was performed [2]. In particular, the authors in [3] analysed the posts on social networks that correlated with postpartum depression and showed that very good results can be obtained in predicting postpartum depression in posts.

Social support in the postpartum period directly affects the birth rate. Through social networks and PPD support groups, opportunities are created for women to share their experiences and receive support [4]. An analysis of changes in the mood of mothers before and after childbirth on Twitter created a model for predicting mood [5]. The possibility of applying machine learning to predict postpartum depression was investigated over the corpus created from Reddit posts in [6]. The characteristic of that most similar research is that they are related to the English language, and non-English social media are quite unexplored. A model for diagnosing postpartum depression via the crowdsourcing platform has been proposed for Serbian, which also includes an automated test for determining the degree of postpartum depression using the Edinburgh scale [7]. This model has been expanded with the detection system of posts that are correlated with postpartum depression by detecting potential users of the crowdsourcing PPD platform on various social networks [8].

3. DATA AND METHODS

For this research, two sources were used: the Reddit forum for posts in English and the ana.rs forum for posts in Serbian (Table 1). Data were collected from both sources according to two criteria:

1. posts related to postpartum depression;
2. posts related to pregnancy and the postpartum period that are not related to postpartum depression.

Data set	Label	Source	URL	Post language
1.	DS1	Reddit	https://www.reddit.com/	English
2.	DS2	ana.rs	https://www.ana.rs/forum/	Serbian

Table 1 – Data sources

The first data set (DS1) consists of Reddit posts from the postpartum depression section (community about: “A non-judgemental place for you to ask for help and vent your frustrations on anything related to issues postpartum, be they hormonal, parental or other mental health issues. PPD, PND, PPA, PPOCD, APD etc. ”) and happy sections (community about: “Too many depressing things on the main page, so post about what makes you warm and fuzzy inside! ") which contain the word “pregnancy ". They were collected using the Pushshift API. 150 posts from the postpartum depression group were singled out, and 150 posts from the happy group.

The second set of data (DS2) is from the forum ana.rs, which is one of the largest women's forums in Serbian. This forum contains the topic of postpartum depression from which 150 posts have been selected.

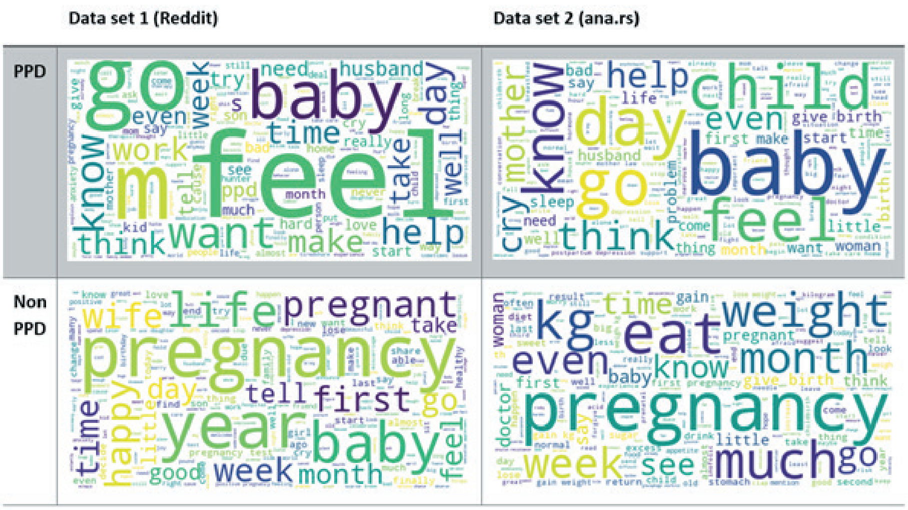

Data set 1 (Reddit)
Data set 2 (ana.rs)
PPD
Non PPD
go feel baby time week want know think help make take day well husband need work
child help day baby go feel know think mother cry even birth sleep husband
wife life pregnant pregnancy year baby tell first week month day time happy good feel
kg eat weight time month know even pregnancy week see much go doctor baby

Figure 1 – Word clouds for contents of PPD post and non-PPD from datasets DS1 and DS2

Also, 150 posts unrelated to postpartum depression were singled out from the sections in “Pregnancy and Weight" and “Pregnancy Chat Corner".

Posts are marked according to the section to which they belong to the group of PPD-related posts and the group of posts not related to PPD. Both sets were manually checked, and posts related to postpartum depression are marked.

3.1. DATA PRE-PROCESSING

Data collected from Reddit are in JSON format, while data from the ana.rs forum are manually processed to CSV format. To achieve data uniformity, both sets went through the preprocessing step.

1. JSON data set (DS1) is transformed into CSV format;
2. The set of data in Serbian (DS2) has been translated into English;
3. Punctuation marks are removed from the data
4. Removals are stop words;
5. Lemmatization

Before processing, it was done using python libraries (*nltk, pandas, gensim, spacy*...). A neural machine translation service[1] that is part of the Azure Cognitive Services family of REST APIs was used for the language translation Serbian post into English. The quality of Microsoft Translator’s machine translation outputs are evaluated using a method called the BLEU score.

BLEU is a measurement of the differences between an automatic translation and one or more human-created reference translations of the same source sentence. In the medical domain that corresponds to the topic of Microsoft Translator’s posts, it has a BLEU score of approximately 50, which is considered high-quality translation.[2]

3.2. TOPIC ANALYSIS

To analyse the contents of the posts, the topic analysis was done for all four groups (two datasets DS1 and DS2 divided by groups if they are related to PPD or not). The topic analysis is performed by using the currently most used LDA (Latent Dirichlet Allocation) topic model developed by David Blei, Andrev Ng i Michael I. Jordan [9]. The python library - *gensim* to construct an LDA model is used.

3.3. CLASSIFICATION METHODS

For the classification of posts into PPD posts and non-PPD posts, methods of classification using Weka tools were performed. Tokenization was performed on the textual data. Sequential minimal optimization (SMO), J48, and RandomForest machine learning methods were applied. Accuracy (Equation 1) was used as the primary measure to gauge the performance of each model. In equation TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

[1] api.cognitive.microsofttranslator.com

[2] microsoft.com/en-us/translator

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 1 - Accuracy

4. RESULTS

This paper aims to analyse the posts related to postpartum depression. Figure 1 contains visual results of the most common words in all four groups of data: PPD and non-PPD posts from DS1 and DS2.

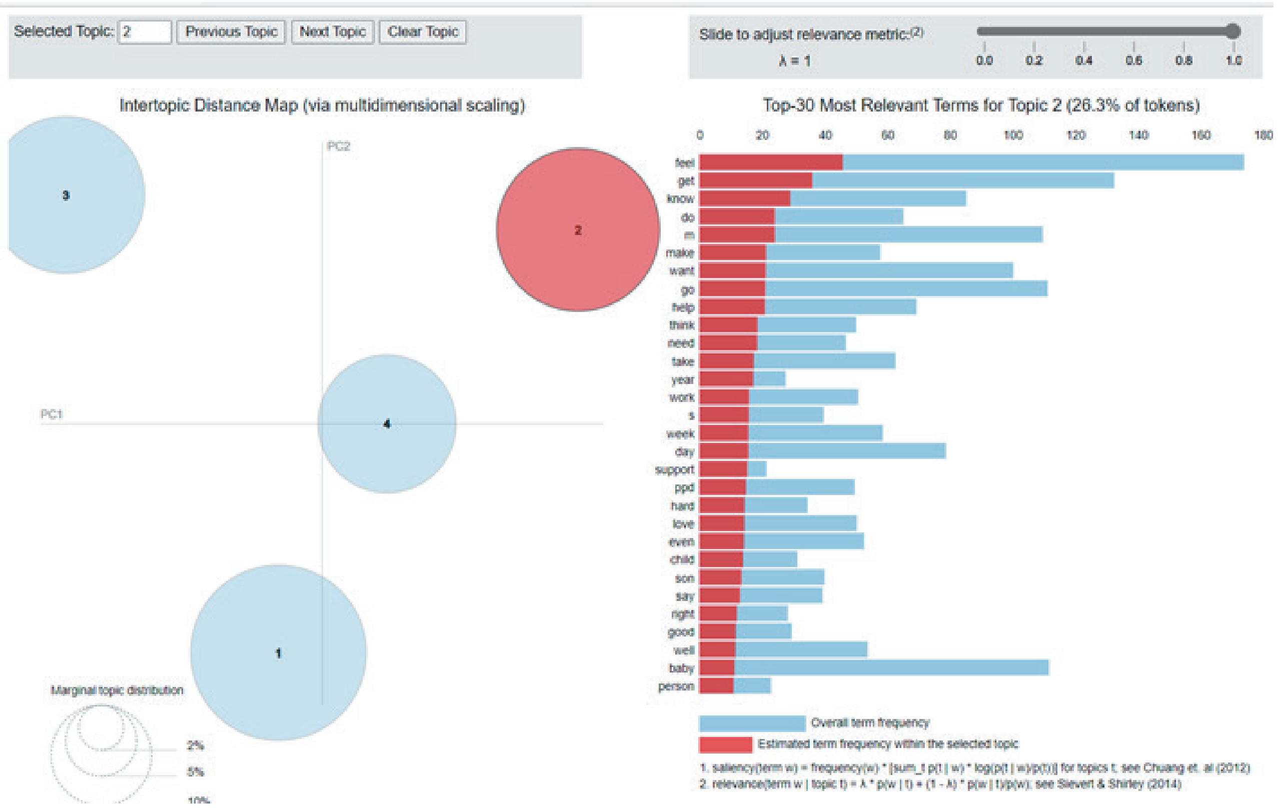
The python *wordcloud* library was used to display the most common words. We can notice that translating from the Serbian data set into English gives a similar word set as in the corresponding word set of English posts from Reddit. The keywords that predominate in PPD-related posts are: feel, baby help think know... While posts that come from groups that have non-PPD refer to: baby, pregnancy...

Using the LDA topic analysis, we attempted to extract topics within datasets. The high degree of coherence (DS1: 0.49, DS2: 0.42) is obtained at high alpha and beta parameters, which indicates a great connection of posts within the set. The best results (high coherence and separation of posts by topic) were obtained for 4 topics (Figure 2, Figure 3).

However, the words that appear in them are still elements of other topics. We obtained similar results at both sets for posts related to postpartum depression. Visualizations of the results show that the terminology is slightly different in posts written in Serbian and posts written in English. English posts contain abbreviations, so they have PPD, while Serbian posts have postpartum depression.

The alpha parameter has values of 0.31 in dataset DS1 and 0.61 for DS1 while the beta parameter has values of 0.9 in both cases. From this we can conclude that each topic will probably contain a mixture of most words.

On a given dataset of posts, classification models were made using methods SMO, J48, and RandomForest (RF). A set of 52 posts from ana.rs from the sections "Postpartum depression", "Pregnancy and Weight" and "Pregnancy Chat Corner" were taken for verification.

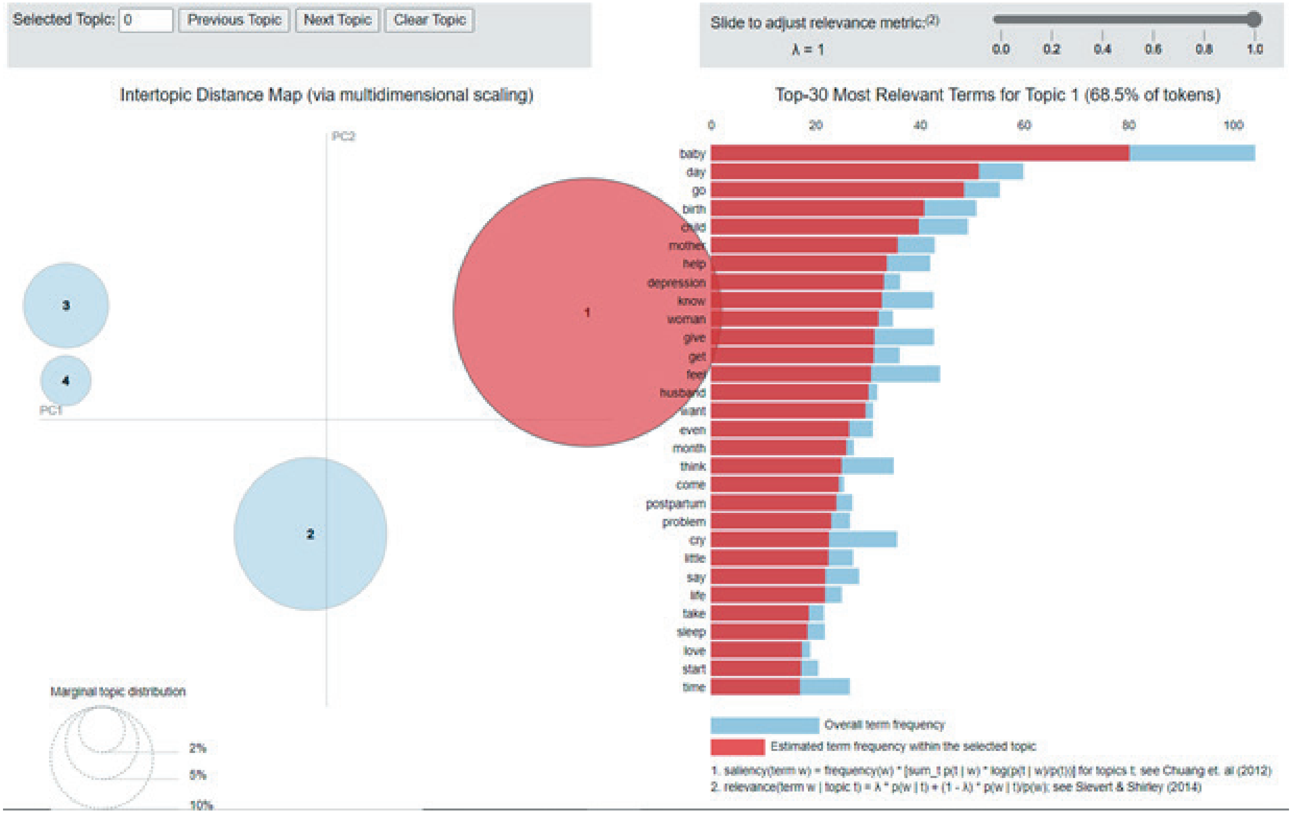


Selected Topic: 2

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 2 – 4 topics from PPD posts (dataset DS1)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3 – 4 topics from PPD posts (dataset DS2)

The first model M1 was made over the posts of the Serbian forum ana.rs (DS2) while the second model M2 was made with posts from both forums (DS1 + DS2).

Model	Dataset	SMO	J48	RF
M1	ds2	92.16	**86.27**	90.20
M2	ds1+ds2	**94.12**	70.59	**92.16**

Table 2 – Classification models

Table 2 shows that by expanding the model with posts from Reddit, better results are obtained. This justifies the fact that for languages with fewer resources, English corpora and tools can alternatively be used. Translation errors affect the results of the classification and should be considered with caution. Anyway, the obtained accuracy is satisfactory and comparable to the results of PPD posts detection in English [3].

5. CONCLUSION

This paper presents an analysis of posts related to postpartum depression from the Serbian and English forums. Data processing was done in English. The datasets showed mutual similarity in content. The obtained classification models on a set of posts from the English and Serbian forums gave satisfactory results. The following research is related to the expansion of the set and the application of more detailed preprocessing methods to more accurately detect posts that are correlated with postpartum depression.

6. ACKNOWLEDGEMENTS

This paper is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under project III44007.

7. REFERENCES

[1] J. L. Cox, J. M. Holden and R. Sagovsky, “Detection of postnatal depression. De-,” *Br J Psychiatry*, vol. 150, pp. 782-786, 1987.

[2] P. A. Cavazos-Rehg, M. J. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj and L. J. Bierut, “A content analysis of depression-related tweets.,” *Computers in human behavior*, vol. 54, pp. 351-357, 2016.

[3] I. Fatima, B. U. D. Abbasi, S. Khan, M. Al-Saeed, H. F. Ahmad and R. Mumtaz, “Prediction of postpartum depression using machine learning techniques from social media text.,” *Expert Systems,* vol. 36, no. 4, p. e12409, 2019.

[4] M. Evans, L. Donelle and L. Hume-Loveland, “Social support and online postpartum depression discussion groups: A content analysis.,” *Patient education and counseling*, vol. 87, no. 3, pp. 405-410.

[5] M. D. Choudhury, S. Counts and E. Horvitz, “Predicting postpartum changes in emotion and behavior via social media.,” in In Proceedings of the SIGCHI conference on human factors in computing systems, 2013.

[6] A. Trifan, D. Semeraro, J. Drake, R. Bukowski and J. L. Oliveira, “Social media mining for postpartum depression prediction,” In *Digital Personalized Health and Medicine*, pp. 1391-1392, 2020.

[7] U. Marovac, A. Ljajić, A. Avdić and A. Fazlagić, “Automation of psychological testing of stressful situations in the Serbian,” in *ICIST 2019 Proceedings*., Kopaonik, 2019.

[8] U. Marovac and A. Avdić, “Detection of postpartum depression-related,” in *The 14th International Symposium on Intelligent Distributed Computing*.

[9] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, no. 3, pp. 993-1022, 2003.