Natural language processing as a tool to identify the Reddit particularities of cancer survivors around the time of diagnosis and remission: A pilot study

Ioana R. Podină[1], Ana-Maria Bucur[1], Diana Todea[1], Liviu Fodor[2], Andreea Luca[1], Liviu P. Dinu[1] and Rareș Boian[2]
[1]University of Bucharest, Romania
[2]Babeș-Bolyai University, Romania
ioana.podina@fpse.unibuc.ro
{ana-maria.bucur,diana.todea}@drd.unibuc.ro

Abstract

In the current study, we analyzed 15297 texts from 39 cancer survivors who posted or commented on Reddit in order to detect the language particularities of cancer survivors from online discourse. We performed a computational linguistic analysis (part-of-speech analysis, emoji detection, sentiment analysis) on submissions around the time of the cancer diagnosis and around the time of remission. We found several significant differences in the texts posted around the time of remission compared to those around the time of diagnosis. Though our results need to be backed up by a higher corpus of data, they do cue to the fact that cancer survivors, around the time of remission, focus more on others, are more active on social media, and do not see the glass as half empty as suggested by the valence of the emojis.

1 Introduction and Related Work

NLP methods have been used to assess and improve treatment outcomes for a plethora of mental health conditions such as depression (Yates et al., 2017; Eichstaedt et al., 2018; Tadesse et al., 2019; Bucur et al., 2021) or anxiety (Ireland and Iserman, 2018; Shen and Rudzicz, 2017). However, they overlooked relevant samples, such as cancer survivors, who have a two/three-time fold risk of developing either one of these conditions relative to non-oncological samples. NLP is gaining ground as a diagnosis and treatment progress tool for several chronic conditions (Koleck et al., 2019), such as cancer (Ribelles et al., 2021). Regarding the cancer community, the behavioral and linguistic patterns of users from an online cancer support community for cancer patients and their caregivers were analyzed and clustered into specific roles such as emotional support provider, story sharer, among others (Yang et al., 2019).

Despite the interest in identifying cancer patients' language particularities, little is known about the language particularities of cancer survivors. By cancer survivors, we understand individuals who entered remission and who finalized their treatment (Little et al., 2000). In the current pilot study, we aimed to use NLP to identify the language particularities of cancer survivors from their Reddit posts around the time of remission compared to those around the time of cancer diagnosis. We assumed that their online behavior would be different around those two key moments of their life regarding their online presence, their emotional state, the use of self-centred references and tenses, and the use of negatively loaded content.

2 Data and Methods

Data To gather our dataset, Reddit users active in remission-related subreddits *r/Remission, r/Cancersurvivors, r/ISurvivedCancer, etc.* were asked to answer a small survey regarding the year of diagnosis, the year of remission, and optionally their Reddit username to extract their publicly available Reddit posts. In addition, several other demographic and survey questions were addressed and will be analyzed on an extended sample size. Out of the approximately 300 individuals contacted, only 60 responded to our questions, and only 39 gave us their consent to extract data from their Reddit posts.

All participants filled in an informed consent before taking part in the survey. Overall, the current work focused on analyzing the social media discourse of cancer survivors around the time of the diagnosis (submissions posted between the diagnosis and remission years, including diagnosis year) and around the time of remission (submissions posted between the time of remission and present time, including remission year). Overall, the pilot study included a total of 15297 submissions (1100 around the time of diagnosis and 14197 around the time of remission, suggesting that cancer survivors were more active in online communities around the time they entered remission), with an average of 413 submissions from each user. The average time between the first to the last submission is 856 days.

Discourse Analysis We analyzed the online discourse, focusing on several features: the frequency of questions, of different POS tags (verbs and pronouns), of emojis and their corresponding sentiment, and polarity. For extracting the POS tags we use the spaCy[1] library. We use the morphological features provided by the library to explore the differences in first, second and third-person pronouns use. To explore verbs usage, we extract the corresponding tenses using The Penn Treebank tagset (Taylor et al., 2003) as follows: VBD and VBN for past tense, VBG, VBZ and VBP for present tense and the MD + VB pair for future tense (Caragea et al., 2018). The frequency of questions is calculated as the number of interrogative sentences divided by the number of all sentences in a submission. We use TextBlob[2] for detecting the polarity of texts, ranging between [-1, 1]. We use the *emoji*[3] library for emoji detection and *emosent*[4] (Kralj Novak et al., 2015) for detecting the corresponding sentiment (positive, neutral or negative).

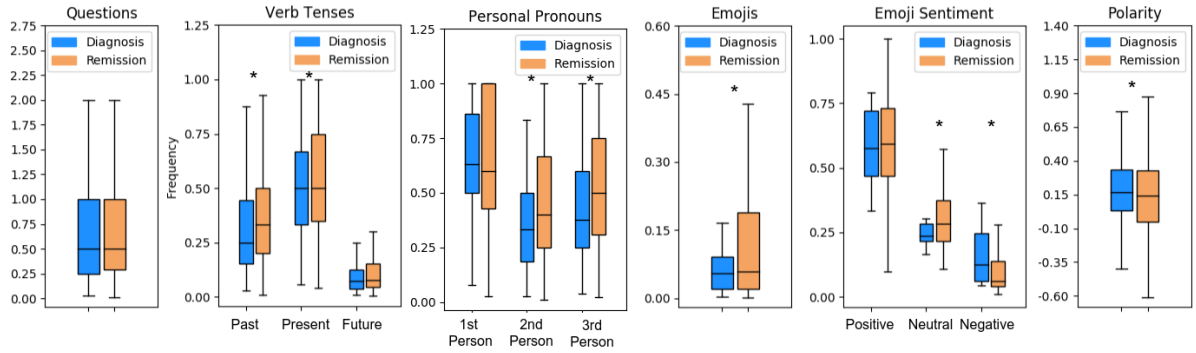

Figure 1: Results of the analysis. *Statistically significant differences (p-value <0.05) using Welch t-test

3 Results

In Figure 1, we present the results of our analysis on the online discourse of cancer survivors, comparing the submissions posted around the time of diagnosis with those around the time of remission. Around the two key moments, we looked into how frequently cancer survivors posted questions on the platform, the tenses they used, the valence of their emojis, and the polarity of their posts, as detailed below.

Questions Regarding the questions addressed online, while the frequency of the questions is not so different in the two periods (around the time of diagnosis and remission), a further analysis is needed to see if the topics of the questions are different.

Part-of-speech Texts around the time of remission contained more past and present tense references. Moreover, around the time of remission, the discourse of cancer survivors tended to revolve more around other persons, with more second and third-person pronouns.

[1] https://spacy.io/
[2] https://textblob.readthedocs.io/
[3] https://github.com/carpedm20/emoji/
[4] https://github.com/FLAIST/emosent-py

Emojis and sentiment analysis The users' posts contained more emojis around the time of remission, with fewer negatively loaded emojis, as opposed to the time of diagnosis. Furthermore, the average polarity of all the posts from our dataset was unexpectedly more positive around the moment of the diagnosis cueing to resilience despite health hurdles that coincide with prior literature on cancer patients, showing positivity around the time of diagnosis (Carrion et al., 2017).

We calculated Cohen's *d* for all the features in Figure 1. Even if some of the differences in the features we analysed are statistically significant, the effect sizes are small ($d < 0.2$). Noteworthy, the results should be interpreted with a pinch of salt. While the extracted texts vary in their properties - as depicted by Figure 1 - depending on the time of diagnosis or remission, more data are needed to conclude that these variations are indeed a reflection of the moment of diagnosis and remission.

4 Conclusion

Though the results of the pilot study need to be backed up by a higher corpus of data, they cue to the fact that cancer survivors focus more on others, are more active on social media, and do not see the glass as half empty in the time around recovery (using fewer negatively loaded emojis) as opposed to the moment of the diagnosis. Interestingly, they display signs of resilience even around the time of diagnosis, as indicated by the positively valenced polarity of their submissions.

A future extension of the pilot study should cover the analyses of other interesting phenomena, such as dysphoric or depressive language manifestations around the time of diagnosis and remission, and cover other social media platforms. Though the implications are distal, a resilient mindset, as expressed by positivity in written language, is an important protective factor against mental health issues (Rude et al., 2004; Ramirez-Esparza et al., 2008). This would explain why not every cancer survivor or cancer patient has depression, in fact, only 15% to 20% display depressive symptoms leaving room for interindividual differences in resilience strategies and coping with illness (Ristevska-Dimitrovska et al., 2015; Riedl and Schuessler, 2021).

5 Acknowledgments

We would like to thank the reviewers for the insightful feedback provided. This work was supported by a Romanian Ministry of Education and Research grant, CNCS-UEFISCDI, project number PN-III-P1-1.1-TE-2019-2140, within PNCDI III.

References

Ana-Maria Bucur, Marcos Zampieri, and Liviu P. Dinu. 2021. An exploratory analysis of the relation between offensive language and mental health. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3600–3606, Online, August. Association for Computational Linguistics.

Cornelia Caragea, Liviu P Dinu, and Bogdan Dumitru. 2018. Exploring optimism and pessimism in twitter using deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 652–658.

Iraida V Carrion, Frances Nedjat-Haiem, Melania Macip-Billbe, and Ryan Black. 2017. "i told myself to stay positive" perceptions of coping among latinos with a cancer diagnosis living in the united states. *American Journal of Hospice and Palliative Medicine®*, 34(3):233–240.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Molly Ireland and Micah Iserman. 2018. Within and between-person differences in language used across anxiety support and neutral Reddit communities. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 182–193, New Orleans, LA, June. Association for Computational Linguistics.

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):364–379.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Miles Little, Emma-Jane Sayers, Kim Paul, and Christopher FC Jordens. 2000. On surviving cancer.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*.

Nuria Ribelles, Jose M Jerez, Pablo Rodriguez-Brazzarola, Begoña Jimenez, Tamara Diaz-Redondo, Hector Mesa, Antonia Marquez, Alfonso Sanchez-Muñoz, Bella Pajares, Francisco Carabantes, et al. 2021. Machine learning and natural language processing (nlp) approach to predict early progression to first-line treatment in real-world hormone receptor-positive (hr+)/her2-negative advanced breast cancer patients. *European Journal of Cancer*, 144:224–231.

David Riedl and Gerhard Schuessler. 2021. Prevalence of depression and cancer–a systematic review. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 67:OA11.

Gordana Ristevska-Dimitrovska, Petar Stefanovski, Snezhana Smichkoska, Marija Raleva, and Beti Dejanova. 2015. Depression and resilience in breast cancer patients. *Open Access Macedonian Journal of Medical Sciences*, 3(4):661.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC, August. Association for Computational Linguistics.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks*, pages 5–22.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.