Establishing control corpora for depression detection in Modern Greek: Methodological insights

Vivian Stamou[1], George Mikros[2], George Markopoulos[1], Spyridoula Varlokosta[1]
[1]National and Kapodistrian University of Athens, Greece
[2]Hamad Bin Khalifa University, Qatar
Panepistimiopoli, Zografou 157 72
Education City, Doha, Qatar 34110
vivianstamou@gmail.com, gmikros@hbku.edu.qa,
{gmarkop, svarlokosta,}@phil.uoa.gr

Abstract

This paper presents a methodological approach for establishing control corpora in the context of depression detection in the Modern Greek language. We discuss various methods used to create control corpora, focusing on the challenge of selecting representative samples from the general population when the target reference is the depressed population. Our approach includes traditional random selection among Twitter users, as well as an innovative method for creating topic-oriented control corpora. Through this study, we provide insights into the development of control corpora, offering valuable considerations for researchers working on similar projects in linguistic analysis and mental health studies. In addition, we identify several dominant topics in the depressed population such as religion, sentiments, health, sleep and digestion, which seem to align with findings consistently reported in the literature.

Keywords: depression detection, control corpora, topic modeling

1. Introduction

NLP research has significantly contributed to depression screening through the development of models for both speech and text applications. The pioneering efforts in depression detection commenced with the groundbreaking work of De Choudhury et al. (2013). Employing crowdsourcing techniques, they identified Twitter users exhibiting symptoms of depression through the CES-D questionnaire (Center for Epidemiological Studies-Depression; Radloff 1977). Their findings revealed distinctive traits among depressed individuals, including reduced social activity, heightened negative emotions, increased self-focus, engagement with medical-related topics, and an elevated expression of religious thoughts. The connection between language and various psychological states was initially articulated by Gottschalk and Gleser (1969) through the Gottschalk method, wherein lexical features extracted from speech data were posited to reflect different psychological dimensions. Building upon this notion, Pennebaker et al. (2003) endeavored to uncover unique linguistic patterns associated with depression. The majority of research investigating the influence of language on depression tends to depend on lexical indicators (both function and content words) rather than larger structures (i.e., sentences), often sourced from dictionaries like Linguistic Inquiry and Word Count (LIWC) (Coppersmith et al., 2014; De Choudhury et al., 2014; Rude et al., 2004; Stirman and Pennebaker, 2001). In addition, alongside lexicon-based methods, topics discussed within textual data have also been employed either independently (Resnik et al., 2015; Tsugawa et al., 2015) or in conjunction with lexical features (Tadesse et al., 2019; Eichstaedt et al., 2018; Resnik et al., 2013).

Social media platforms have played a crucial role in examining mental health disorders, serving as virtual communities that encompass two dimensions: communication (i.e., the interaction among users) and social status indication (i.e., users' self-representation). Furthermore, a benefit of these platforms is the ability to collect metadata information such as socio-demographic details (e.g., age, gender), time span, location, and user-network information. This enables a more comprehensive understanding of users within the virtual space, while also facilitating the tracking process in case of a disease outbreak (Li and Cardie, 2013; Schmidt, 2012).

Typically the data collection methods for depression detection in social media platforms involve four approaches (Guntuku and et al., 2017). In the first approach, which is based on crowd-sourced surveys, users fill out a depression questionnaire and then share their Facebook or Twitter content (De Choudhury et al., 2013; Tsugawa et al., 2015). This method enables the assessment of their mental health status, as the questionnaire-derived information helps determine whether they belong to the depressed or to the control population. The second approach, self-reported diagnoses, target users who are identified through self-declarations (e.g., 'I was diagnosed with depression'). The latter

Proceedings of the RaPID-5 @LREC-COLING 2024, pages 68–76

was introduced in the 2015 Computational Linguistics and Clinical Psychology (CLPsych) workshop[1]. A third approach, described as the participation at specific blog communities, involves data collection from users registered in online forums (such as Reddit[2]; De Choudhury et al. 2016). Finally, data can be directly extracted from social media platforms based on keywords ("data which contain words drawn from a specific vocabulary"), and subsequently post-processed by human experts following specific annotation guidelines (Prieto et al., 2014). In this study, we opted for the second approach in order to target users experiencing depression. Moreover, we chose not to employ crowdsourcing to collect candidate users, considering the potential challenges posed by the Greek Twittersphere, and rather followed an automated method.These challenges include the difficulty of reaching a large crowd due to concerns about privacy, anonymity, and the stigma associated with disclosing mental health issues. These factors could deter individuals from openly participating in crowdsourced data collection efforts (Naslund et al., 2015). In addition, users who voluntarily participate may systematically differ from those who do not, ultimately impacting the generalizability of findings.

The paper is organized as follows: in Section 2 we review previous studies related to techniques utilized for constructing control corpora (i.e., corpora representing the normal population). Section 3 outlines our methodology for compiling the depression corpus, including also the establishment of control corpora through two methods: random selection and consideration of topics identified in the corpus of depressed users. Specifically, we present the methodological approach for generating the topic-oriented control corpus and the results of topic modeling using various pretrained models both monolingual and multilingual. Finally, in Section 4, we provide a summary of the key findings.

2. Previous Work

Various techniques have been employed to create a control corpus (CC) of non-depressed individuals. Chancellor and De Choudhury (2020) identify five ways of constructing a control corpus sample: (i) CC is checked and evaluated in order to ensure it does not contain people having a mental disorder (De Choudhury et al., 2013; Guan et al., 2015) ; (ii) CC is created based on the application of random selection among social media users, thus the process does not guarantee the inclusion of people with mental health issues (Mitchell et al., 2015; Coppersmith et al., 2014, 2015b); (iii) CC data is collected considering specific criteria which are indicative of the absence of mental health issues. For instance, users' interests and participation in communities related to mental health topics or selection of users who had never used in their posts related terms to depression (Shen et al. 2013; Yates et al. 2017); (iv) CC is derived according to matching criteria such as demographic and behavioral properties (i.e., age and gender; Coppersmith et al. (2015b); Landeiro Dos Reis and Culotta (2015) and the selection of a specific time span (Li et al., 2019); and (v) CC dataset is different from the original dataset (Orabi et al. 2018; Soldaini et al. 2018).

Furthermore, in order to exclude true positive cases from the data, sampling techniques have been utilized to focus on particular social media users. Rafail (2018) underscores the importance of sampling as a significant yet frequently overlooked aspect of managing databases containing social media content. He further proposes a typology, categorizing populations into three distinct types based on the methodology used in constructing the database. These categories include unbounded populations (i.e., no restrictions applied), semibounded populations, and bounded populations. More specifically, semibounded populations are also divided into user-restricted by means of selecting users who fulfill certain criteria and topic-restricted, when these are drawn around a particular topic. Nevertheless, the amalgamation of both methodologies results in bounded populations.

3. Corpora compilation

With respect to the construction of the depression corpus, we employed a combination of user- and topic-bounded sampling techniques. Initially, we initiated the process by searching for a specific keyword, which in our case refers to the declarative statement indicating depression. Subsequently, we selectively sampled content or history exclusively from the identified target users (i.e., individuals experiencing depression). Sampling strategies for collecting data within the social media landscape are typically classified as either probability/random-based or non-probability-based. In forming the control corpus, we prioritize random sampling methods to select from our target user pool. Consequently, the initial phase involves gathering random data from the Greek Twitter, as elaborated in subsection 3.2.

3.1. The depression corpus

Data was collected by searching tweets in which users explicitly acknowledged that they had been

[1] The CLPsych Shared Task (Coppersmith et al., 2015a) focused on the implementation of Machine Learning methods to differentiate between Twitter users with depression and users with Post Traumatic Stress Disorder (PTSD).

[2] https://www.reddit.com/

diagnosed with depression. Self-disclosure diagnosis is a common technique used to collect data in such cases (Jagfeld et al., 2021; Shin et al., 2020; Jamil et al., 2017; Coppersmith et al., 2014). For search purposes, we implemented the Twitter API[3] through which it was possible to go back to the history of each user. In total 2,500 tweets were extracted, which were then checked manually in order to avoid cases of humor or references coming from articles in health newsites. The final number of real self-statements of depression was 110 and belonged to 51 Twitter users. We collected tweets for the time period between September 2018 and June 2020[4]. The final corpus size reached up to 659,189 tweets by downloading the full user's history.

3.2. Randomly sampled control corpus

Our methodological approach for compiling the random control corpus relies on language-specific data (i.e., tweets in the Greek language) and further aligns with the methodology introduced by Bergsma et al. (2012). In their endeavor to extract language-specific content, they employ two primary methods. Firstly, they gather data from users identified as sources, who simultaneously serve as 'hubs' with a significant number of followers and who tweet in the target language. These sources are continually updated by collecting their followers and retrieving their tweets. Secondly, they identify users who tweet in the language of interest through the 'geotagging' method, allowing them to query tweets based on specific latitude and longitude coordinates.

Based on Bergsma et al. (2012), we created our control dataset by searching for data limited to the Greek language and exploiting geolocation information. To access Twitter's API, we used the Tweepy Python library[5]. Considering that retrieving Twitter content typically necessitates a textual reference like a term or hashtag, and given that most tools employ multifaceted queries, the inclusion of geolocation information proved crucial in refining the selection of tweets. There are different techniques for approaching language identification (LI), such as the implementation of specific tools (i.e., langid.py; Lui and Baldwin 2012, or compact language detector (CLD2)[6]). However, the implementation of such tools requires more effort given that irrelevant language data should be cleaned. Therefore, we targeted language-specific content by querying Twitter with the Where on Earth ID (WOEID) code and utilized the method GET trends/place provided by the Twitter API. The GET trends/place function allows developers to retrieve the top 50 trending topics for a specific location. Therefore, once a list of geolocated trends was obtained, it became feasible to gather data containing those hashtags (trends), thereby enabling access to the users generating such content. Initially, the total number of users was 502.

We subsequently expanded the user population with possible candidate users by searching their network and retrieving their followers. The latter was possible via the GET followers/list endpoint. The possible candidate users were limited to the most popular ones (i.e., users with many followers) by including only those having over 1000 followers. Among the methods used to measure user popularity is the follower-rank measure, which indicates the number of followers a user has (Cha et al., 2010). We prioritized popular users due to our expectation of a higher likelihood of tweet volume. Out of a total of 800,000 Twitter users, we selected 100,000 users randomly, ensuring each user had an equal probability of inclusion, thereby reducing bias in the data selection process. Following this chance-oriented approach, we obtained a final list of users and collected their tweet history. Ultimately, we randomly sampled a corpus of 100,000 tweets from a total of 27 users.

3.3. Topic-oriented control corpus

The second control corpus was derived considering the topics of discussion in the depression corpus. For this reason, we applied a topic modeling analysis in the depression corpus. Topic models are employed to unveil latent themes (i.e., topics) or subjects within collections of text, without prior information. These topics are defined as sets of words that collectively represent specific domains, such as education or health. Several previous studies that have considered depression identification in social media have aimed to utilize topics as a means of discovering the most dominant themes in depressed language (Resnik et al., 2013, 2015; Tsugawa et al., 2015; Eichstaedt et al., 2018; Tadesse et al., 2019).

In order to derive the topics from the depression corpus we utilized the BERTopic library (Grootendorst, 2022), which is flexible in allowing the selection of various embedding models. BERTopic utilizes a deep neural network architecture, namely BERT (Bidirectional Encoder Representations, Devlin et al. 2019), which has been trained on a big amount of textual data and which can be further specialized to downstream tasks, such as document classification, sentiment analysis etc.

[3] https://developer.twitter.com/en/docs/twitter-api

[4] Twitter's social platform, now renamed X, has undergone rebranding and adjusted limitations on data retrieval. However, it is important to note that our data collection occurred prior to these changes.

[5] https://github.com/tweepy/tweepy

[6] https://github.com/CLD2Owners/cld2

BERTopic firstly generates document embeddings via BERT and subsequently clusters topics into semantically similar clusters through two steps: (i) employing Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of embeddings, and (ii) utilizing Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to cluster the reduced embeddings (McInnes et al., 2018, 2017). Finally, the model generates topics and extracts class-specific words to create keywords for each topic.

We employed several available pretrained models, accessible through Hugging Face[7], both monolingual and multilingual, to generate the sentence embeddings, without any prior corpus preprocessing, as seen in Table 1. The pretrained models include both monolingual, namely Greek-BERT-Base-Uncased-V1 (Koutsikakis et al., 2020), GreekSocialBERT (Alexandridis et al., 2021), the RoBERTa Greek base model[8], as well as multilingual models like stsb-xlm-r-greek-transfer developed by the Hellenic Army Academy (SSE) and the Technical University of Crete (TUC), all-MiniLM-L6-v2 and distiluse-base-multilingual-cased-v2 (Reimers and Gurevych, 2019). Subsequently, we opted to narrow down the number of topics to the top 100 most significant ones and computed coherence scores for each model by calculating the C_v measure[9]. This measure quantifies the distance among words within a topic, as provided by the gensim library (Řehůřek and Sojka, 2010).The advantage of C_v measure lies in its ability to handle indirect similarities between words. In particular, it also addresses cases where certain words should be grouped together within a topic despite their infrequent co-occurrence (Röder et al., 2015).

Next, we evaluated the performance of different models to determine which one yielded the most favorable outcomes for the resulting topics. The outcome is presented in Table 1, where a score closer to 1 indicates higher coherence. The highest coherence score, namely 0.54, was achieved by `roberta-el-news`, which is a model trained on 8 million news articles. However, we decided to manually inspect the results of each model. Manual evaluation was conducted by a Greek native speaker and the `stsb-xlm-r-greek-transfer` model was selected as best, which also can account for mixed language. This model has the capability to handle cases with mixed language, typically found in social media, because of the incorporation of the transfer learning approach (i.e., trained on parallel EN-EL sentence pairs). This design ensures the integration of vocabulary from English language as well, enabling us to extract topics such as `μωρή_μωρό_μωράκι_baby`/ `silly_baby_little baby_baby`. In Figure 1 below, we provide the similarity matrix of the selected model generated by calculating the cosine similarities for the topic embeddings. In particular, the Figure depicts how specific topics relate to each other. Denser blue areas are indicative of a high similarity score. For instance, topic 86 <καλή ενέργεια θετική>/ <good positive energy> is related to topic 61 <ζωή ευτυχία ζωής>/<life happiness of life> with a score 0.848.

Models	**Number of topics**	**C_v**
`bert-base-greek-uncased-v1`	100	0.5099
`greeksocialbert-base-greek-uncased-v1`	100	0.4465
`stsb-xlm-r-greek-transfer`	100	0.3960
`roberta-el-news`	100	0.5466
`all-MiniLM-L6-v2`	100	0.4598
`distiluse-base-multilingual-cased-v2`	100	0.4364

Table 1: Embedding models and their coherence scores.

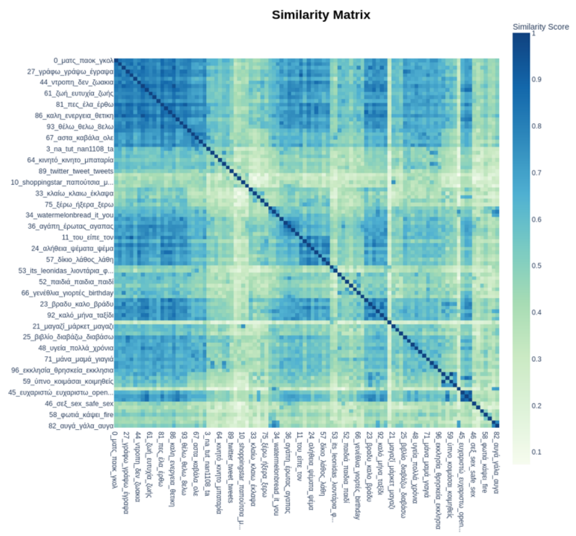

Figure 1: Similarity matrix of the stsb-xlm-r-greek-transfer model.

Following that, ChatGPT3.5 (OpenAI, 2023) was employed in a prompt-based manner to cluster the top 100 topics into 20 clusters. The prompt used to derive the clusters was the following one: "I will provide you with a list of words. Could you please arrange them into 20 clusters?". Although Chat-

[7] `https://huggingface.co/`

[8] cvcio/roberta-el-news

[9] Only the C_v measure from the gensim library yields reasonable scores, while other measures consistently produce negative scores, raising concerns about result reliability with respect to the metric implementation. Further discussion and cautionary notes can be found here: `https://github.com/dice-group/Palmetto/issues/12`

Topic	Count	Name-Translation
-1	421844	να_και_το_δεν to_and_the_not
0	9162	ματς_παοκ_γκολ_ομαδα - match_paok_goal_team
1	5672	the_and_to_is
2	4810	ευρώ_λεφτά_τα_για - euro_money_the_for
3	4654	na_tut_nan1108_ta_to
4	3537	survivorgr_survivorpanoramagr
5	3475	youtube_via_μέσω_χρήστη - youtube_via_via_user
6	3115	χιούμορ_γελάω_γέλιο_χιουμορ - humor_laugh_laughter_humor

Table 2: Topic counts and names.

GPT's output is not perfect, an automated way was provided to rapidly cluster a large set of complex topics. The clusters targeted the following domains: (1) sports, (2) money, (3) social media, (4) alcohol, (5) shopping, (6) animals, (7) time, (8) religion, (9) sentiment, (10) stores, (11) love, (12) health, (13) truth and lies, (14) leisure activities, (15) relationships (i.e., wedding, family), (16) sleep, (17) digestion, (18) sex, (19) dream, life, and (20) elections. Subsequently, it was possible to retrieve tweets by looking at a representative keyword for each cluster. In this way, a topic-oriented control corpus of 9 million tweets was collected for the time period between January 2018 and June 2023.

Basic preprocessing was applied to all the corpora which includes the removal of duplicates, html tags, emojis, universal resource locator (URL) and the "@" indicator that denotes usernames. Detailed statistics for all corpora are included in Table 3. NA stands for not applicable since this control corpus is not created based on specific users but considering keywords/topics instead.

Data set	Users	Total Tweets	Mean Tweets	SD
DC	51	659,189	10.919	33.8236
CC_random	27	100,000	127.541	61.5508
CC_topic-oriented	NA	600,000	111.99	58.47

Table 3: Dataset statistics.

4. Conclusion

In this work, we discuss several methodological approaches for datasets sourced from the social media platform of Twitter in our effort to create a dataset that differentiates between depressed and non-depressed users in the Modern Greek language. We narrow down our selection to two distinct strategies: randomly sampling a corpus from prominent Twitter users and constructing a control corpus aligned with the topics relevant to individuals experiencing depression. Interestingly, some of the topics detected in the depression corpus have been reported in many studies to be highly correlated with depression (De Choudhury et al., 2013; Resnik et al., 2013; Eichstaedt et al., 2018; Tadesse et al., 2019). In particular, these topics refer to terms related to religion, sentiment, health, sleep and digestion. Both datasets are created as adjuncts to, rather than substitutes for, clinicians. We anticipate that the methodology presented will provide valuable support for mental health professionals. Currently, we are experimenting with both machine learning and deep learning techniques to distinguish between the two populations based on specific language indicators.

Additionally, relying on topic modeling techniques to construct corpora based on similar topics allows for a more focused examination of the linguistic content of users. As a result, the comparison between two population types is not entirely random but rather constrained or associated with a specific topic. This approach offers the advantage of potentially achieving a more nuanced differentiation based on language indicators, thereby highlighting subtle differences in expression. For example, it enables the investigation of how users expressing depression differ from those who do not within a given topic.

5. Copyrights

6. Acknowledgements

This work is part of the first author's doctoral thesis. «The implementation of the doctoral thesis was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities».

7. Limitations

We acknowledge certain limitations in our study. Firstly, regarding the creation of the randomly sampled control corpus, it is important to discuss the constraints of relying primarily on geo-tagged data. Previous research has demonstrated that geo-tagged tweets may exhibit demographic biases (Karami et al., 2021). Additionally, relying heavily on popular accounts could potentially skew the control sample. Ideally, the inclusion of demographic information would enable a more comprehensive examination of differences between the two populations. Moreover, it is crucial to acknowledge the potential bias in terms of fluency, especially considering the association between depression and alogia, typically referred as poverty of content of speech (Kaplan and Sadock, 2008). Alogia is a symptom of depression expressed as reduced speech, which is attributed to a disruption in the thought process. While prioritizing popular users may enhance the richness of our dataset in terms of volume, it is important to highlight the potential impact on the linguistic quality of the content.

8. Ethics statement

This work complies with the ACL Ethics Policy.[10]. Twitter is a public platform where users share information openly. For this reason, it is essential to respect the privacy and anonymity of individuals who may be mentioned or involved in the data collected, especially in the context of mental health data. To ensure anonymity, we have removed any direct identifiers such as usernames and any other personally identifiable information from the dataset. An approval from the Institution's Ethics Committee is not required for the following reasons. As Twitter data are publically available, users are aware of the fact that their content can be seen and analyzed by anyone (Kamocki et al., 2022; Mikal et al., 2016). In addition, data are distributed in compliance with Twitter company policy and terms of service[11], while access to both the depression and the control corpora will be granted exclusively to researchers who consent to adhere to ethical guidelines. These guidelines encompass restrictions against contacting or attempting to deanonymize any of the users. Furthermore, in the application of GPT-3.5 was restricted solely to organizing a larger volume of topics into the top-20 most prominent ones. As a result, we did not touch upon sensitive domains, but rather focused on this specific task.

To gain access to the dataset, please contact the authors directly, ensuring compliance with ethical guidelines outlined in this section.

[10] https://www.aclweb.org/portal/content/acl-code-ethics

[11] https://twitter.com/en/tos#intlTerms

9. Bibliographical References

Georgios Alexandridis, Iraklis Varlamis, Konstantinos Korovesis, George Caridakis, and Panagiotis Tsantilas. 2021. A survey on sentiment analysis and opinion mining in greek social media. *Information*, 12(8).

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74, Montréal, Canada. Association for Computational Linguistics.

M Cha, H Haddadi, F Benevenuto, and K Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 10–17.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine*, 3:43.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015a. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Glen Coppersmith, Cassandra Harman, and Mark Dredze. 2015b. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, Eric Horvitz, and Andrea Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the ACM Confer-*

ence on Computer Supported Cooperative Work (CSCW), pages 625–637.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*, pages 128–137.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes C Eichstaedt, Ryan 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 of the United States of America*, 115(44):11203–11208.

Louis A Gottschalk and Goldine C Gleser. 1969. *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Lin Guan, Bibo Hao, Qijin Cheng, Paul S.F. Yip, and Tingshao Zhu. 2015. Identifying chinese microblog users with high suicide probability using internet-based profile and linguistic features: Classification model. *JMIR Mental Health*, 2(2):e17.

Sharath C Guntuku and et al. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven Jones. 2021. Understanding who uses Reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 1–14, Online. Association for Computational Linguistics.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Pawel Kamocki, Vanessa Hannesschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Lindén, and Andrius Puksas. 2022. Legal issues related to the use of twitter data in language research. pages 68–75.

Harold I. Kaplan and Benjamin J. Sadock. 2008. Chapter 4 signs and symptoms in psychiatry. In *Kaplan and Sadock's Concise Textbook of Clinical Psychiatry*, page 29. Wolters Kluwer/Lippincott Williams & Wilkins.

Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi. 2021. Analysis of geotagging behavior: Do geotagged users represent the twitter population? *ISPRS International Journal of Geo-Information*, 10(6).

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

Vinicius Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health via twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Genghao Li, Bing Li, Langlin Huang, and Sibing Hou. 2019. An automatic construction of depressing-domain lexicon based on microblogs (preprint).

Jiwei Li and Claire Cardie. 2013. Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Jude Mikal, Sam Hurst, and Mike Conway. 2016. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC Medical Ethics*, 17(1):22.

Margaret Mitchell, Glen Coppersmith, and Kristy Hollingshead, editors. 2015. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Association for Computational Linguistics, Denver, Colorado, USA.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, Gregory J McHugo, and Stephen J Bartels. 2015. Crowdsourcing for conducting randomized trials of internet delivered interventions in people with serious mental illness: A systematic review. *Contemporary Clinical Trials*, 44:77–88.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

OpenAI. 2023. ChatGPT. Software.

Haya Orabi, Rafael A Calvo, David N Milne, and Mohsin Husain. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Victor M Prieto, Sergio Matos, Miguel Álvarez, Fidel Cacheda, and Jose L Oliveira. 2014. Twitter: A good place to detect health conditions. *PloS one*, 9(1):e86191.

Lenore S Radloff. 1977. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1:385–401.

Patrick Rafail. 2018. Nonprobability sampling and twitter: Strategies for semibounded and bounded populations. *Social Science Computer Review*, 36(2):195–211.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Philip Resnik, Megan E Armstrong, Livia Claudino, and Thai Nguyen. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353, Seattle, Washington, USA. Association for Computational Linguistics.

Philip Resnik, Philip Resnik, Megan E Armstrong, Livia Claudino, and Thai Nguyen. 2015. Beyond lda: Exploring supervised topic modeling for depression-related language in twitter. In *CLPsych@HLT-NAACL*, volume 30.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Sabine Rude, Eva-Maria Gortner, and James W Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition Emotion*, 18:1121–1133.

Charles W Schmidt. 2012. Trending now: Using social media to predict and track disease outbreaks. *Environmental Health Perspectives*, 120(1):a30–a33.

Yu-Chun Shen, Tsung-Ting Kuo, I-Ning Yeh, Tzu-Ting Chen, and shou-de Lin. 2013. Exploiting temporal information in a two-stage classification framework for content-based depression detection. volume 7818, pages 276–288.

Dongwook Shin, Ki-Jae Lee, Tolu Adeluwa, and Jaehyung Hur. 2020. Machine learning-based predictive modeling of postpartum depression. *Journal of Clinical Medicine*, 9(9):2899.

Bib Soldaini, T. Walsh, A. Cohan, J. Han, and N. Goharian. 2018. Helping or hurting? predicting changes in users' risk of self-harm through online community interactions. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology*, pages 194–203. Association for Computational Linguistics.

Shannon Stirman and James Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4):517–522.

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

Sho Tsugawa, Yoshihiko Kikuchi, Fumiya Kishino, Kohei Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196, New York, NY, USA. Association for Computing Machinery.

Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067. ACM.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.