Building a corpus on Eating Disorders from TikTok: challenges and opportunities

Melissa Donati, Ludovica Polidori, Paola Vernillo, Gloria Gagliardi
Alma Mater Studiorum - University of Bologna, Italy
{melissa.donati,ludovica.polidori}@studio.unibo.it
{paola.vernillo,gloria.gagliardi}@unibo.it

Abstract

We present two synchronic corpora of Eating Disorders (ED) related discourse on Social Media. PAC (i.e., ProAna/Anorexia Corpus) and RAC (i.e., Recovery from Ana/Anorexia Corpus) resources focus on the contents posted on TikTok, respectively, by communities promoting anorectic behavior and users sharing experiences concerning the process of recovery from their ED. We report on the corpus statistics and creation process, focusing specifically on the methodological issues raised by this novel Social Media platform.

1 Introduction

It was only 20 years ago that one of the darkest sides of eating disorders (ED) was revealed through the proliferation of websites, blogs, and social networks, in which a growing number of adolescents and young adults started sharing information about their eating experiences with like-minded users. Among these pro-ED communities, researchers and clinicians showed particular concern for pro-ana (i.e., “pro-anorexia”) groups, i.e., web-based communities of anorexic (or aspiring anorexic) individuals engaged in the promotion of their eating disorder (Boero and Pascoe, 2012). Interestingly, one of the most horrific and dangerous aspects of pro-ana groups is that anorexia nervosa (AN) is not presented as a psychiatric disorder associated with pathological body image dissatisfaction (Williams and First, 2013), but more as a way of living with its own rules and rituals to be respected. While over the last years, much has been done to prevent the circulation of pro-ED content on social media (e.g., TikTok’s adoption of measures to obscure harmful contents: Marsh (2020)), a new but specular phenomenon recently took the toll, that is, the spread of pro-recovery accounts of individuals who are in the process of healing from an ED and are willing to share their eating experience to help other online users (Greene et al., 2023). From a linguistic perspective, research on ED has been very limited and became an object of study only in recent years (Bates, 2015; Knapton, 2013; Lyons et al., 2006; Skårderud, 2007a; Skårderud, 2007b; Wolf et al., 2013) as opposed to other psychopathologies, such as schizophrenia (Bambini et al., 2016; De Boer et al., 2020), personality disorder (Arntz et al., 2012), and depression (Bernard et al., 2016; Brockmeyer et al., 2015; Ramirez-Esparza et al., 2008; Zimmermann et al., 2017). This already problematic picture has been further compromised by the inhomogeneous representation of linguistic data in the literature, where the majority of studies have been dedicated to the linguistic profiling of ED-affected individuals in a Germanic language (Cuteri et al., 2021). This paper represents a small step towards the reversal of this tendency but a crucial part of two larger projects (Metaphan[1] and RaAM project 2022[2]) aiming at identifying, by the adoption of different NLP techniques and tools, potential lexical and semantic patterns in anorectic individuals. To this end, in the current research, we show the data collection process (i.e. oral and written productions) from ED communities on TikTok, currently representing the most widely used social media among young people and adolescents, namely the population groups at greater risk for ED. In the following paragraphs, we give a brief overview of the literature on the topic (Section 2), then we describe the process of creating the corpus and discuss the methodological issues that were met (Section 3) and to conclude we provide few insights for future works (Section 4).

[1] https://site.unibo.it/metaphan/en

[2] https://site.unibo.it/metaphan/en/connected-research-activities

2 Related Works

In recent years, we have witnessed exponential growth in the use of Social Media (SM), especially by adolescents and young people. The community-building nature and the interactive dynamics of these platforms, as well as the less direct way of communicating, encourage users to openly discuss a wide variety of topics (Lenhart et al., 2010). In turn, this makes available huge amount of data that can be used for different purposes (e.g. extract actionable patterns, form conclusions about users, conduct research, etc.). For this reason, Social Media Mining (SMM), i.e., the process of extracting big data from SM, now constitutes a well-established methodology to collect large samples of data in different research areas (Gundecha and Liu, 2012). This approach has proved particularly fruitful for collecting data on ED as people suffering from these disorders seem to overcome the self-protective nature of their ED to engage in ED-related discourse with online users sharing similar experiences (Kenny et al., 2020). Indeed, in the last decade, many studies have used different SM platforms as a source of data to analyze EDs (Lukač and others, 2011; Mullany et al., 2016; Moessner et al., 2018; Bohrer et al., 2020; Kenny et al., 2020; Herrick et al., 2021; Jordan et al., 2021; González-Nuevo et al., 2021; Minadeo and Pope, 2022). However, the state-of-the-art on ED-discourse on SM currently presents two main limitations: i) the majority of the analysis was carried out on small datasets built ad-hoc for the purpose of the work (with the only exception of Donati and Strapparava (2022)), and ii) they mostly focused on the English language. As a matter of fact, in the Italian framework there have been very little research on the representation of ED on SM, and that little was mostly focused on Anorexia-Nervosa and did not target ED in general (Richichi et al., 2018; Bragazzi et al., 2019; Gagliardi, 2021).

3 Corpus Creation: Methodological Issues

Against this background and intending to fill this gap, we created a collection of English and Italian ED-related data that could be used for different types of analysis and research. We selected TikTok as a source of data as it currently represents the most widely used SM, especially among young people and adolescents, namely the at-risk population for ED (Sherman, 2020).

To achieve this goal, we first needed to define the nature and characteristics of the corpus itself. As far as the linguistic features are concerned, our corpus is specialized (i.e., is focused on the topic of ED discourse on TikTok), synchronic (i.e., refers to a specific point in time that is the moment the data were downloaded), and targets both written and spoken language (TikTok videos contain spoken and/or written text). We did not set *a priori* a target dimension to be reached, because this feature is totally dependent upon the possibility of extracting the data automatically (Section 3.1). Conversely, following the common practice in the domain of SMM, we assumed that 'there is no data like more data' and intended to download as many videos as possible. To maximize the corpus representativity, we tried to balance the sample with respect to the types of videos being collected but we could not do so concerning the users' gender, because for both corpora the vast majority of profiles were of female individuals (see Section 3.3 for more details). The target population consisted of those profiles that identify themselves in one of the two following categories: i) supporters of anorectic behaviors (for the English PAC corpus); ii) witnesses and motivators for the recovery process (for the Italian RAC corpus). Such profiles were identified based on the linguistic and non-linguistic (i.e., emojis) information present in their profile bio. The selection criteria will be presented in Section 3.1, prior to the description of the data collection process and the discussion of the related issues that were encountered.

3.1 Data Collection

As explained above, the selection criteria adopted to identify the target profiles was based on the information present in the profiles' bio. However, to track the target profiles, we needed to start from a list of ED-related hashtags that could lead us to such profiles via a keyword-based search. The hashtags that were used herein were generated both by brainstorming and by exploring the platform for a couple of weeks, noting down the most popular trends and the most widely used hashtags (see Table 1 for an overview). Following this hashtag-driven search, we noticed that there was very little -if any- pro-Ana content produced in Italian, that is why for this type of ED-related content we decided to collect a small sample of English data. On the other hand, we found quite some profiles representing the ED-recovery community. Among these profiles, we selected those having at least 10k followers (some of them exceed 2M followers) and at least 10 ED-related posts, so that we could maximize the chance of gathering interesting and relevant linguistic information. Having identified the target profiles, we used the ED-related hashtags to conduct a within-profile research to select only the ED-related videos in each profile in order to extract them.

At this point, the next step consisted of extracting the identified ED-related videos from the selected profiles. For the sake of time and efficiency, we wanted to download the data automatically. However, differently from other popular SM, TikTok has not yet released any official API that can be used by researchers and developers to automate the process of accessing and extracting the data. In addition, even if unofficial APIs exist, they get outdated almost immediately after their release because TikTok is constantly updating the anti-bot system preventing automatic access from the same IP. To get around this, we looked for a reliable and cost-effective proxy provider for TikTok scraping, but we could not find any viable solution.

Therefore we decided to proceed with the manual downloading of the data. The main drawback of this way of proceeding is that due to time and resource constraints we could not collect a very large number of videos (see Table 1). On the bright side, however, the manual downloading allowed us to i) enhance the content filtering process and ii) notice that TikTok videos have different formatting styles that might be worth distinguishing not only to ease the ensuing transcription process but also to conduct separate content analysis and compare the different results. Based on our observations about the different formatting styles, we grouped the TikTok videos into 4 subcorpora: 1)*Speech-onl*y videos: in which the user was talking in the absence of background music and/or written text; 2) *Playback*: in which the user lip-sync over a song or an extract from a movie or tv shows; 3) *Text-only*: in which there is neither background music nor the users themselves speaking, but only written text superposed on the video; and 4) *Mixed*: in which the above-mentioned features are present in various combinations.

Pro-Ana hashtags	**Pro-Recovery hashtags**
#weightloss (w3ightl0ss) #unhealthyweightloss (+ lexical.variations) #kpop	#dcarecovery (dcar3covery) #dca #dcaitalia #fiocchettolilla #dcafighting

Table 1: List of pro-ana and pro-Recovery hashtags that were used to search for TikTok profile that share ED-related content

3.2 Transcription

Organizing the videos into 4 categories was particularly useful for the transcription phase as it allowed to adopt different strategies and techniques based on the input characteristics. As for the downloading phase, although we intended to automatize the transcription process as much as possible, the high complexity of the data has, in some cases, made human intervention necessary. For speech-only and playback videos automatic transcription was performed using the Google Web Speech API, which is easily accessible through the SpeechRecognition Library (Zhang et al., 2017). To assess the quality of the automatic transcription, a random sample of videos (n=10) for each category were extracted, transcribed manually and then compared with the machine-based transcription. For speech-only videos, a high agreement score was obtained between human and machine transcription (>90%) which confirmed the viability of the method adopted. Conversely, playback videos emerged as more problematic, thus manual correction was needed because both singing and the music accompaniment adversely impacted on intelligibility. Automatic transcription was also attempted for text-only videos by means of Optical Character Recognition (OCR) using the Tesseract OCR engine (Ooms, 2023), but we obtained poor results due to the high visual complexity of the input data, more specifically to the extreme variability of font type, size, and color, the lack of adequate contrast with the background, the non-hierarchical spatial organization of texts, and the presence of non-textual graphical elements (e.g., lexical variations of words, where letters are substituted by numbers or emojis to prevent the platform's censorship and filtering system from blocking the content as potentially harmful, e.g., 'starving' written replacing star with the corresponding emojis, or ‘disorder’ written as ‘d1s0rder’). The same issue, boosted to the maximum, was observed with mixed videos, where speech, music, and written text were mingled. Therefore, for these two categories of videos, we could only perform the transcription manually.

3.3 Corpus Statistics

In Table 2, we reported an overview of the statistics for the two corpora in terms of number of videos, number of words, and number of users from whose profiles the data were extracted.

	PAC	**RAC**
n videos	250	1000
n words	13169	116261
n users	14 (all F)	27 (26 F, 1 M)

Table 2: Statistics for the two corpora

4 Conclusion and Future Works

The aim of this work was twofold: on the one hand, we wanted to present two corpora on ED, the English pro-Ana corpus (PAC) and the Italian pro-Recovery corpus (RAC), that were both built by extracting data from the popular SM TikTok; on the other, we wanted to discuss some methodological issues related to building a corpus using this platform as a source of data. More specifically, we pointed out that the absence of an official API does not allow the automatic extraction of the videos and requires manual work, which is highly time-consuming and does not allow to collect a very large sample of data. This, in turn, might impede the application of more complex computational analysis and limit the generalizability of the results. In addition, we raised the issue related to the transcription of the videos to text. In this case, implementing automatic approaches is not always feasible because of the extreme visual complexity and variability of TikTok videos. Given the highly interactive nature of this SM and its unprecedented success, we believe that TikTok constitutes an extremely interesting source of linguistic and non-linguistic data that could be used to analyze other complex social and psychological phenomena and we hope that this work paves the way for further research in this direction.

CRediT authorship contribution statement

MD Conceptualization, Methodology, Software, Data Curation (i.e., download, automatic transcription, annotation), writing (§2,3,4) **LP** Data Curation (i.e., manual transcription) **PV** Conceptualization, Data Curation (i.e., download), Writing (§1) **GG** Supervision, Funding acquisition.

Funding

This work was partially fundend by the RaAM Association (project “How about metaphors for dinner? A digest of metaphorical conceptualizations in pro-ana communities”) and the University of Bologna (AlmaIdea 2022 - “MetaphAN” project).

References

Arnoud Arntz, Lisa D Hawke, Lotte Bamelis, Philip Spinhoven, and Marc L Molendijk. 2012. Changes in natural language use as an indicator of psychotherapeutic change in personality disorders. *Behaviour research and therapy*, 50(3):191–202.

Valentina Bambini, Giorgio Arcara, Margherita Bechi, Mariachiara Buonocore, Roberto Cavallaro, and Marta Bosia. 2016. The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Comprehensive psychiatry*, 71:106–120.

Carolina Figueras Bates. 2015. “i am a waste of breath, of space, of time” metaphors of self in a pro-anorexia group. *Qualitative Health Research*, 25(2):189–204.

Jared D Bernard, Jenna L Baddeley, Benjamin F Rodriguez, and Philip A Burke. 2016. Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *Journal of Language and Social Psychology*, 35(3):317–326.

Brittany K Bohrer, Una Foye, and Tom Jewell. 2020. Recovery as a process: Exploring definitions of recovery in the context of eating-disorder-related social media forums. *International Journal of Eating Disorders*, 53(8):1219–1223.

Nicola Luigi Bragazzi, Giulia Prasso, Tania Simona Re, Riccardo Zerbetto, and Giovanni Del Puente. 2019. A reliability and content analysis of italian language anorexia nervosa-related websites. *Risk management and healthcare policy*, pages 145–151.

Timo Brockmeyer, Johannes Zimmermann, Dominika Kulessa, Martin Hautzinger, Hinrich Bents, Hans-Christoph Friederich, Wolfgang Herzog, and Matthias Backenstrass. 2015. Me, myself, and i: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in psychology*, 6:1564.

Vittoria Cuteri, Giulia Minori, Gloria Gagliardi, Fabio Tamburini, Elisabetta Malaspina, Paola Gualandi, Francesca Rossi, Milena Moscano, Valentina Francia, and Antonia Parmeggiani. 2021. Linguistic feature of anorexia nervosa: a prospective case–control pilot study. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, pages 1–9.

JN De Boer, M Van Hoogdalem, RCW Mandl, J Brummelman, AE Voppel, MJH Begemann, E Van Dellen, FNK Wijnen, and IEC Sommer. 2020. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *npj Schizophrenia*, 6(1):10.

Melissa Donati and Carlo Strapparava. 2022. CorEDs: A corpus on eating disorders. In *Proceedings of the RaPID Workshop - Resources and Process-Ing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - within the 13th Language Resources and Evaluation Conference*, pages 80–85, Marseille, France, June. European Language Resources Association.

Gloria Gagliardi. 2021. “odio tutto ciò, voglio le ossa”: Una prima indagine sulle caratteristiche linguistiche delle pagine social pro-ana in lingua italiana. *Italiano LinguaDue*, 13(1):520–536.

Covadonga González-Nuevo, Marcelino Cuesta, and José Muñiz. 2021. Concern about appearance on instagram and facebook: Measurement and links with eating disorders. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 15(2).

Amanda K Greene, Hannah N Norling, Lisa M Brownstone, Elana K Maloul, Caity Roe, and Sarah Moody. 2023. Visions of recovery: a cross-diagnostic examination of eating disorder pro-recovery communities on tiktok. *Journal of Eating Disorders*, 11(1):109.

Pritam Gundecha and Huan Liu. 2012. Mining social media: a brief introduction. *New directions in informatics, optimization, logistics, and production*, pages 1–17.

Shannon SC Herrick, Laura Hallward, and Lindsay R Duncan. 2021. “this is just how i cope”: An inductive thematic analysis of eating disorder recovery content created and shared on tiktok using# edrecovery. *International journal of eating disorders*, 54(4):516–526.

G Lladò Jordan, MDC Dìaz Garcìa, B Lozano Dìez, P Mediavilla Sànchez, JA Gòmez Del Barrio, and R Ayesa-Arriola. 2021. Facebook as a pro-ana and pro-mia resource. *European Psychiatry*, 64(S1):S703–S703.

Therese E Kenny, Sarah L Boyle, and Stephen P Lewis. 2020. # recovery: Understanding recovery from the lens of recovery-focused blogs posted by individuals with lived experience. *International Journal of Eating Disorders*, 53(8):1234–1243.

Olivia Knapton. 2013. Pro-anorexia: Extensions of ingrained concepts. *Discourse & Society*, 24(4):461–477.

Amanda Lenhart, Kristen Purcell, Aaron Smith, and Kathryn Zickuhr. 2010. Social media & mobile internet use among teens and young adults. millennials. *Pew internet & American life project.*

Morana Lukač et al. 2011. Down to the bone: A corpus-based critical discourse analysis of pro-eating disorder blogs. *Jezikoslovlje*, 12(2):187–209.

Elizabeth J Lyons, Matthias R Mehl, and James W Pennebaker. 2006. Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation. *Journal of psychosomatic research*, 60(3):253–256.

Sarah Marsh. 2020. Tiktok investigating videos promoting starvation and anorexia. *The Guardian*, 7.

Marisa Minadeo and Lizzy Pope. 2022. Weight-normative messaging predominates on tiktok—a qualitative content analysis. *Plos one*, 17(11):e0267997.

Markus Moessner, Johannes Feldhege, Markus Wolf, and Stephanie Bauer. 2018. Analyzing big data in social media: Text and network analyses of an eating disorder forum. *International Journal of Eating Disorders*, 51(7):656–667.

Louise Mullany, Catherine Smith, Kevin Harvey, and Svenja Adolphs. 2016. ‘am i anorexic?’weight, eating and discourses of the body in online adolescent health communication. *Communication & medicine*, 12(2-3).

Jeroen Ooms, 2023. *tesseract: Open Source OCR Engine*. https://docs.ropensci.org/tesseract/ (website) https://github.com/ropensci/tesseract (devel).

Nairan Ramirez-Esparza, Cindy Chung, Ewa Kacewic, and James Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. In *Proceedings of the international AAAI conference on web and social media*, volume 2, pages 102–108.

Veronica Richichi, Alessandro Chinello, Francesca Parma, Luigi Enrico Zappa, Elvis Mazzoni, and Fiorella Monti. 2018. Anoressia nervosa e internet. uno studio sui blog pro-ana in italia. *Psicologia clinica dello sviluppo*, 22(3):499–514.

Alex Sherman. 2020. Tiktok reveals detailed user numbers for the first time. *Retrieved October*, 2:2020.

Finn Skårderud. 2007a. Eating one's words, part ii: The embodied mind and reflective function in anorexia nervosa—theory. *European Eating Disorders Review: The Professional Journal of the Eating Disorders Association*, 15(4):243–252.

Finn Skårderud. 2007b. Eating one's words, part i:'concretised metaphors' and reflective function in anorexia nervosa—an interview study. *European Eating Disorders Review: The Professional Journal of the Eating Disorders Association*, 15(3):163–174.

Janet BW Williams and Michael First. 2013. Diagnostic and statistical manual of mental disorders. In *Encyclopedia of social work*.

Markus Wolf, Florian Theis, and Hans Kordy. 2013. Language use in eating disorder blogs: Psychological implications of social online activity. *Journal of Language and Social Psychology*, 32(2):212–226.

Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, Cesar Laurent Yoshua Bengio, and Aaron Courville. 2017. Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*.

Johannes Zimmermann, Timo Brockmeyer, Matthias Hunn, Henning Schauenburg, and Markus Wolf. 2017. First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical psychology & psychotherapy*, 24(2):384–391.