Supporting the Annotation Experience Through CorEx and Word Mover's Distance

Stefania Pecòre
School of Electrical Engineering and Computer Science, University of Ottawa, Canada

Abstract

Online communities can be used to promote destructive behaviours, as in pro-Eating Disorder (ED) communities. Research needs annotated data to study these phenomena. Even though many platforms have already moderated this type of content, Twitter has not, and it can still be used for research purposes. In this paper, we unveiled emojis, words, and uncommon linguistic patterns within the ED Twitter community by using the Correlation Explanation (CorEx) algorithm on unstructured and non-annotated data to retrieve the topics. Then we annotated the dataset following these topics. We analysed then the use of CorEx and Word Mover's Distance to retrieve automatically similar new sentences and augment the annotated dataset.

2012 ACM Subject Classification Applied computing → Document management and text processing; Applied computing → Annotation

Keywords and phrases topic retrieval, annotation, eating disorders, natural language processing

Digital Object Identifier 10.4230/OASIcs.LDK.2021.12

Funding IT12441 MITACS and SafeToNet Canada

Acknowledgements We thank MITACS and SafeToNet Canada for their generous funding. In addition to this, we thank the University of Ottawa and the supervisor of the project, Professor Diana Inkpen, for their support.

1 Introduction

Online social platforms provide an easy way to share ideas, opinions, information, and personal messages. Research suggests that online communities are a support tool for recovery and promotion of self care and well-being [21]. At the same time, these platforms may be used to enhance and promote destructive behaviours as in pro-Eating Disorder communities (pro-ED groups). Eating disorders such as *anorexia nervosa, binge eating disorder, and bulimia nervosa* are recognized as mental disorders in standard medical manuals (ICD-10[1] and DSM-5[2]). The exact etiology of eating disorders remains unclear [35, 13] and they are a real concern due to the highest mortality rate of any mental illness, affecting various ethnic groups [28], males and females [15], any age range [16], with a highest peak during teen age [3]. During the last 10 years the research community has been analysing pro-ED groups using different platforms: Instagram [9, 8], Tumblr [14], Flickr [46], Reddit [32, 38], YouTube [39], Twitter [1, 45]. The analyses have been carried out according to different points of view: social media moderation [7, 9], relation between pro-ED users and ED content [1, 34], contrast between similar – "thinspiration" and "fitspiration" [41], online ED content analysis [5, 48, 4, 19, 40], pro-ED users' identity perception [2, 20], ED markers [33], multimodal classification [7], and early detection of anorexia signs [42]. ED and mental illnesses have also been the main focus of recent workshops such as CLEF E-risk and

[1] https://www.who.int/classifications/icd/icdonlineversions/en/
[2] https://www.psychiatry.org/psychiatrists/practice/dsm
[3] https://www.eatingdisorderhope.com/information/statistics-studies

CLPsych [27, 26, 25, 12]. According to Twitter rules and policies[4], it is not possible to promote or encourage self-harming behaviours (including eating disorders). However, Twitter has not banned or restricted the access to any specific pro-ED related hashtags and content, allowing us use this data for research purposes.

The main contributions of this paper are: (a) a new use of Correlation Explanation (CorEx) algorithm [44] to retrieve topics, emojis, and contextual foreign words in English tweets of native and non-native English Eating Disorder communities, (b) the use of the Word Mover's Distance (WMD) model [24] to annotate similar sentences and assist the annotators. We aim to create a tool to assist annotators in their annotation task, by providing a way to semi-automatically increment the number of annotated sentences, even in a complex context such as the ED communities. To the best of our knowledge, the amount of models annotating 10 years of tweets in emojis and non-common word patterns related to Eating Disorders (ED) with the use of CorEx to extract ED topics [48, 17] and Word Mover's Distance to assist annotators is limited at this time. We believe that this work could be of interest for the research community also in other domains, where topic extraction and annotation are involved.

We will describe our dataset (Section 2), then we will present the CorEx algorithm (Section 3) and the reasons behind the choice of this algorithm instead of others frequently used such as LDA. We will explore how we used CorEx to retrieve documents correlated with the topics (Section 4) and, consequently, why and how we decided to manually annotate our dataset for ED aspects (Section 5). Finally, we will describe our approach with Word Mover's Distance to assist annotators in data annotation tasks (Section 6), and we will identify the limitations of this work (Section 7), followed by conclusions and future work.

1.1 Ethical Considerations

This work uses public tweets from 2009 to 2019. No personally identifiable information (location, photos, names) was used in this study, nor was included in any of our algorithms. We did not interact with the subjects of this study, and since the data is public, we did not need institutional review board approval. The annotators were given anonymized data.

2 General pro Eating Disorder (pro-ED) Twitter dataset

2.1 Data Collection

In order to create our initial dataset, we collected tweets by using known ED tags [8, 14, 6] through a library[5] preserving Unicode emojis. From the seed tags, we retrieved both related posts and ED hashtags (a partial list is shown in table 1). We found low frequency hashtags that were not related to the ED and others that – without a context – seemed not directly related to ED (#casuloana, #whale, #borboletana) and their presence became more clear during the annotation phase. The datasets are in English and available under request – due to NDA reasons.

[4] https://help.twitter.com/en/rules-and-policies/glorifying-self-harm
[5] https://pypi.org/project/GetOldTweets3/

Table 1 Examples of hashtags found after retrieving the data.

Keywords Found	No. of # in the dataset	Keywords	No. of # in the dataset
#thinspo	78,244	#skinny	7,446
#proana	38,692	#ana	6,213
#thinspiration	10,471	#weightloss	5,322

2.2 Data preprocessing

The initial dataset was composed of 106,793 tweets dated from 2009 to 2019. During the normalization phase we transformed words in lowercase and removed most of the non-ED related tweets – e.g. tweets that had more than 80% of the content about link referrals. We also removed duplicates and punctuation (except the symbol # to preserve the hashtags). We applied the fastText Language Identification tool [23, 22] as a filter to avoid non-English sentences. We applied a basic anonymization filter by replacing tagged user, with the corresponding label USER, numbers with the label NUM, websites with the label URL, common cities with the label LOC. After the normalization and anonymization phases, our dataset had 87,957 entries.

From the analysis of our dataset (Table 2 & Table 3) we noticed a low lexical diversity: there are only 67,296 types[6] over 1,150,508 tokens in total. Moreover, the pro-ED community on Twitter seems to prefer writing on average short tweets: less than 11 words per tweet and 41–60 characters (Figure 1). We expected that after Twitter's characters doubling in 2017, the most recent tweets would have been longer. This was the case only for 1% of the tweets.

Table 2 Dataset lexical analysis.

No of tweets	No of tokens	No of types	Type/Token ratio
87,957	1,150,508	67,296	5,85%

Table 3 Average, Median and Standard Deviation of Words and Characters per tweet.

Distribution of WORDS per tweet		Distribution of CHARACTERS per tweet	
AVG	11.8	**AVG**	75.2
MDN	10	**MEDIAN**	65
STDEV	7.04	**STDEV**	42.3

2.3 Emojis

In order to have emojis present later in our tests, we decided to translate them from Unicode to their description. The description has been taken from the CLDR Short Name information repository[7]. For example:

- the *Unicode U+1F600* corresponds to the description *grinning face*,
- the *Unicode U+1F605* corresponds to the description *grinning face with sweat*.

We noticed a low frequency in conjunction with a low diversity in the use of the emojis: only 11.75% of the lines showed one or more emojis and only 4.75% of emoji types have been used.

[6] `https://plato.stanford.edu/entries/types-tokens`
[7] `https://Unicode.org/emoji/charts/full-emoji-list.html`

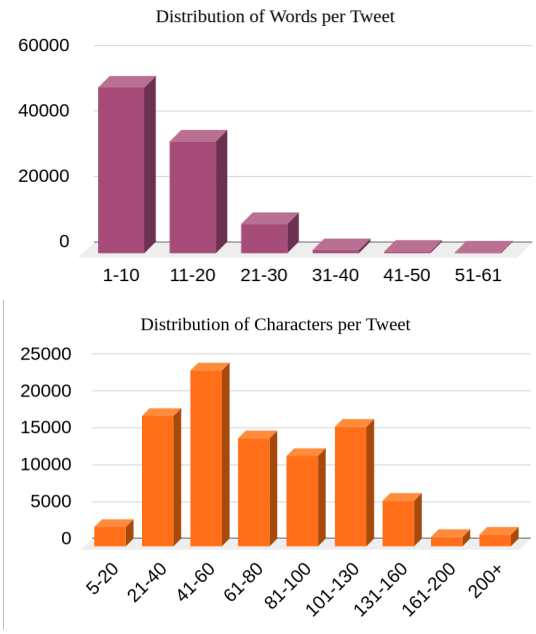

Figure 1 Distribution of words and characters per tweet.

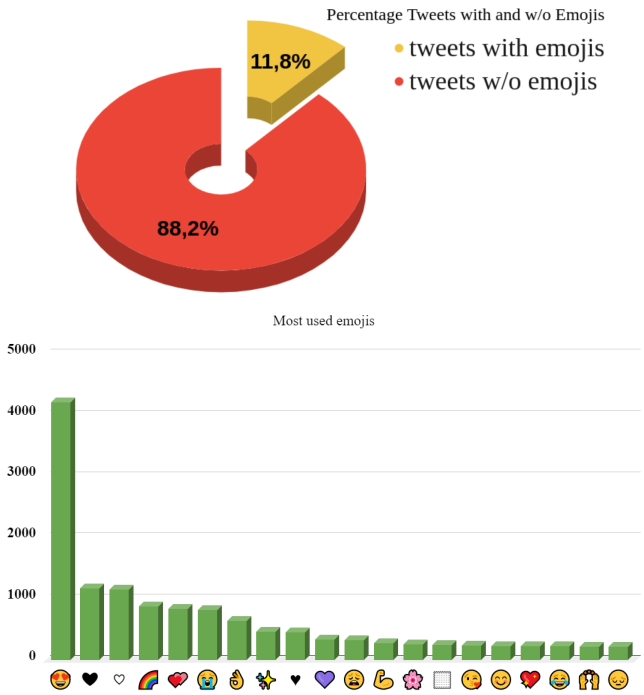

Figure 2 Distribution of emojis in the dataset.

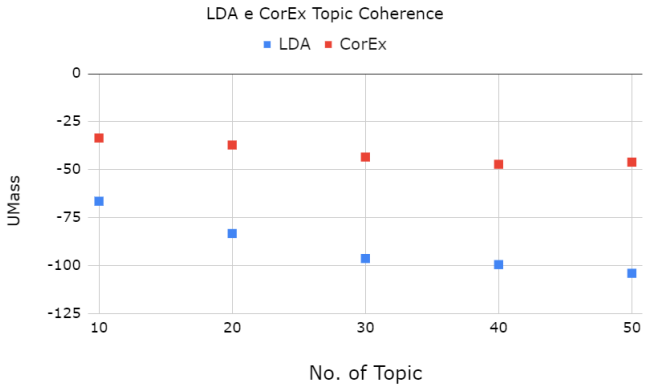

Figure 3 Comparison of CorEx to LDA with respect to topic coherence on ED Dataset.

We noticed that only 15% of the most used emojis seem to convey a negative sentiment. We remarked that positive polarity emojis don't pair always with an overall positive content: the first most used emoji, *face with heart shaped eyes*, has been used for the appreciation of emaciated bodies. Even though emojis will be present on our experiments and results, they are only a part of our work, analysis, and findings on this type of communication.

3 Finding Eating Disorder related topics through CorEx

3.1 Why did we choose CorEx?

Research has already explored the use of topic modeling to assist document annotation [43, 37, 47, 11, 48]. For our experiments we needed a model able to extract topics in unstructured and low-diversity data (see subsection: 2.2). According to the authors of CorEx [17], the model should work better than LDA models [3], since it maximizes the mutual information between words and topics without any assumption on how documents are generated [17]. Plus, according to the authors, CorEx is a discriminative model that works very well with minimal domain knowledge, which is the case when pre-annotated data is not present. In order to test whether CorEx works better than LDA also for our specific dataset, we tested it against LDA in detecting semantic topic quality as in [18]. As the authors wrote, CorEx does not explicitly attempt to learn a generative model and, traditional measures such as perplexity are not appropriate for model comparison against LDA. Furthermore, it is well-known that perplexity and held-out log-likelihood do not necessarily correlate with human evaluation of semantic topic quality [10, p. 6]. For this reason, we measured the semantic topic quality using Mimno et al.'s [31] UMass topic coherence score, which correlates with human judgments.

We used the same dataset for both models and we varied the number of topics. We ran the model 30 times for each time we changed the number of topics. We tested the models using 10, 20, 30, 40, 50 topics. For both models, we noticed the tendency to obtain worse topic coherence when the number of topics increased – see Figure 3. Each dot is the average of 30 runs. We suppose that the low lexical variety is due to the small number of topics discussed. For both models, the optimal number of topics seems to be 10. When we compared CorEx to LDA, we noticed that CorEx outperformed LDA in terms of topic coherence. For this reason, we decided to use CorEx to identify and describe the latent topics from our dataset. The final number of chosen topics is 10.

Once the number of topics has been chosen, two experts manually reviewed the words of the topics learned by CorEx to re-verify the coherence and the content. The categories, their description, and some examples are shown in Table 4. The results of the evaluation confirmed the ED topics and keywords that have been described in other ED related studies [8, 14, 6].

3.2 Results from CorEx application

During the manual review of the extracted words, we found both emojis and special words connected to the topics. We decided purposely not to remove them, because we believe that emojis, as well as special words, could indicate a presence of eating disorder content.

We carried out an in-depth analysis of special words that are relevant to pro-ED communities. It seems interesting to highlight that, although they seem not relevant, they are uncommon contextual words. Following Table 4, here are some interesting findings:

1. **foreign language keywords** – from row ED OTHER LANGUAGES: even though we removed non-English sentences, there are some sentences that are written in English, with hashtags or a partial content in another language. We were able then to capture hidden contents also in other languages. An example is the hashtag #waniliowemleko ["*vanilla milk*" in Polish] that has been found also in some pro-ED websites[8] citing a popular drink within the ED community, that seems to be used as a social drink with few calories. We noticed that these words are usually associated with other ED English relevant keywords, such as *#skinny* and *#diet*;
2. **acronyms** – from DIET row: we noticed that the model was able to capture words that may appear of difficult interpretation without prior knowledge and a context. For examples: *NF* for "*no food*", *OMAD* for "*One Meal A day*", *ABCDIET* for "*Ana Boot Camp diet*", *NT* for "*no thanks*" (usually linked to food offer as in "*dinner? NT*");
3. **celebrities** – in row SPORT: there are references to some YouTube celebrities ("*Lena Snow*", "*Chloe Ting*", "*Alexis Ren*") who stream weekly workouts;
4. **ED slang**: the model also unveiled other words that may not seem significant without prior knowledge, but they refer to encrypted community slang. We refer to words such as "*borboletana*" (from "*borboleta*", butterfly in Portuguese, a shared symbol of pro-ED and recovery communities, and "**AN**orexi**A**"), "*casuloana*" ("*casul of Ana*", from "casual" of ana that in Internet slang means newbie of ana),"*rexy*" ("ano**RE**xia + se**XY**") and *#skinnylegend* (which represents skinny photoshopped celebrities' body);
5. **relevant ED emojis**: we found that the emojis issued from the model are in line with the words associated and they:
   a. **may reinforce the meaning of the word** – from the row CONSEQUENCES: the emojis [mouth], [nauseated face], [face with open mouth vomiting] seem to be properly associated with words such as *"purge"*, *"binge"*, *"puke"* that appear in the dataset;
   b. **may publicly manifest user's gender**, such as [female sign] in GENERAL ED
   c. **may express sarcasm**, such as [face upside down] found in WEIGHT row. This emoji is present in sentences like "*yesterday I ate like a normal person cus I was with my family and woke up 1.2 lbs heavier [face upside down] fuck*" and "*I'm going on a binge. Can't wait to purge [face upside down] I'm such a fat ass.*")
   d. **may express more than words**, such as [dizzy symbol] in SPORT row.

[8] `https://www.wattpad.com/334686866-sad-skinny-girl-guide-eating-out-starbucks`

Table 4 Some examples of retrieved topics with their descriptions, related words, and emojis found (the ratio between content words and hashtags displayed here is not representative of the distribution of this dataset.)

Topic Category	Description	Words	Emojis
ED words in other languages	not English words related to ED	#abwtbs #bslyw #waniliowemleko #caspfb38 #samajl	NA
GENERAL ED	General References to pro-ED content	#proana #atypicalanorexia #slimthickspo #casuloana #edtwt #edtwitter #edproblems	[face_with_tears_of_joy] [female_sign] [butterfly] [person_shrugging] [face_with_pleading_eyes]
WEIGHT	Everything related to the person's weight and weight management	lose weight lbs gain pound #goal #weightcheck cw ugw sw #bodycheck	[face_upside_down]
BODY REPRESENTATION	how they judge themselves compared to another body not in a measurable way	#butterfly #borboletana #dysmorphia #fattie #fatpig #rexy	[butterfly] [broken_heart]
BODY DESCRIPTION	how they describe a body in a neutral manner	bone collar collarbone hip #hipbone thigh gap flat waist cm #fat	NA
SELF REPRESENTATION	How they judge themselves in a not measurable way	bitch cow #whale stupid whore ugly #ass cunt dumb #lazy #pathetic	[whale]
HARM	Everything related to self harm and risky consequences for the person	#depression #anxiety #selfharm #selfhate #cutting #deadinside laxative pills	[face_with_medical_mask] [knife
COMMUNITY	Interactions within the pro-ED group	#meanspocoach #sweetspo #bonespo #nicespo #skinnylegend #malespo rexy	[smiling_face_with_heart] [smiling_face_with_smiling_eyes] [white_heart]
DIET	Everything related to diet and calories	#stopeating NT NF #donteat #foodfears #OMAD fasting #abcdiet calorie intake	[raised_fist]
SPORT	Everything related to sport and activities to burn calories	workout lena snow chloe ting alexis ren gym routine #itsallfine	[thumbs_up_sign] [dizzy_symbol] [skull_and_crossbones] [flex_biceps]
CONSEQUENCES	person's action impacts on him/her life	stomach hurt pain growl grumble purge force parent haven	[mouth] [nauseated_face] [face_with_open_mouth_vomiting]

4 Automatic Retrieval of documents through CorEx

We then used an internal function of CorEx to retrieve the documents with the topics discovered before, because we believe that using CorEx to retrieve topics and also documents at the same time could make the annotation process faster and smoother. We retrieved the most probable documents per topic. According to the paper [18], CorEx estimates the logarithmic probability of a document belonging to a topic given that document's words. In order to evaluate this process, we retrieved the first 100 most probable documents for the topics *Weight, Body Representation, Self Representation, Harm, Consequences, Community, Sport, Diet, Body Description, and General ED*. We then created guidelines explaining the type of topic described by CorEx using some examples, and we asked two native English speakers to evaluate whether a sentence belonged to a certain topic or not. Except for the topic *General ED*, we noticed that the results were not satisfying (see Table 5). Since we were not able to retrieve the documents directly from the estimation of the probability of CorEx, we decided to use another algorithm to discover, given a seed of annotated documents this time, other documents similar to the annotated ones. The chosen algorithm is Word Mover's Distance (WMD) [24]. We chose this algorithm because it targets both semantic and syntactic information to calculate similarity between text documents. It is designed to overcome the synonym problem: since similar words should have similar vectors, WMD can calculate the distance even when there are no common words.

In order to have the same type of dataset to evaluate this method, we chose to annotate a part of the pro-ED dataset. Then, we ran the experiments against this dataset in a controlled environment.

Table 5 F1-score for 100 most-probable documents belonging to the topic.

	Weight	Body Representation	Self Representation	Harm	Consequences	Community	Sport	Diet	Body Description	General
F1	0.0178	0.1291	0.0159	0.0275	0.0315	0.0676	0.0334	0.0564	0.1326	0.6412

5 Annotation scheme and examples

The annotation scheme is composed of ten categories. We decided to remove the *ED words in other languages* category for simplification since the tweets were not completely written in English. Here are some examples for each category:

1. **Body Description (BD)**: "I **am** skinny"
2. **Body Representation (BR)**: "I **want** to be skinny **as** her"
3. **Community (COM)**: "I love my #anasisters"
4. **Consequences (CON)**: "Finally today i eat And a feel a little bit bad"
5. **Diet (D)**: "tomorrow I'm going to start fasting again"
6. **General ED (G)**: "Being thin and not eating are signs of true will power and success #proana"
7. **Harm (H)**: "I don't like laxatives but it's time"
8. **Self Representation (SR)**: "I am perfect", "I woke up and I was still UGLY thanks for nothing"
9. **Sport (S)**: "Insanity kicked my butt What a good workout"
10. **Weight (W)**: "Losing weight is good, gaining weight is bad"

We would like to make three specific categories explicit, as we did before the beginning of the annotation phase with the annotators, as they may be cause of confusion:

- **Body description** is applied to each sentence where there is a statement about a body (where the person describes the body). Examples are: *"she's so skinny"*, *"I am skinny"*, *"look at her collarbones!"*
- **Body representation** is applied to each sentence where the people refer to themselves by means of a comparison with other people. Examples are: *"I want to be skinny as her"*, *"I want her legs"*
- **Self representation** is applied to each sentence where people judge themselves using not measurable and imaginary words. Examples are: *"I'm ugly"*, *"I'm perfect"*, *"I would like to be graceful as a butterfly"*

5.1 Human Annotation process and guidelines

The annotation has been done via PigeonXT[9] by two English native speakers. The annotation was done at the sentence level to identify the topics. In total, 3,064 sentences (Table 6) have been annotated. Even if a tweet can be as short as the average of 11 words, we assumed that it was possible to find more than one category per tweet. However, we decided to take into consideration only one category at a time, considering the overall content complexity. At the

[9] `https://github.com/dennisbakhuis/pigeonXT`

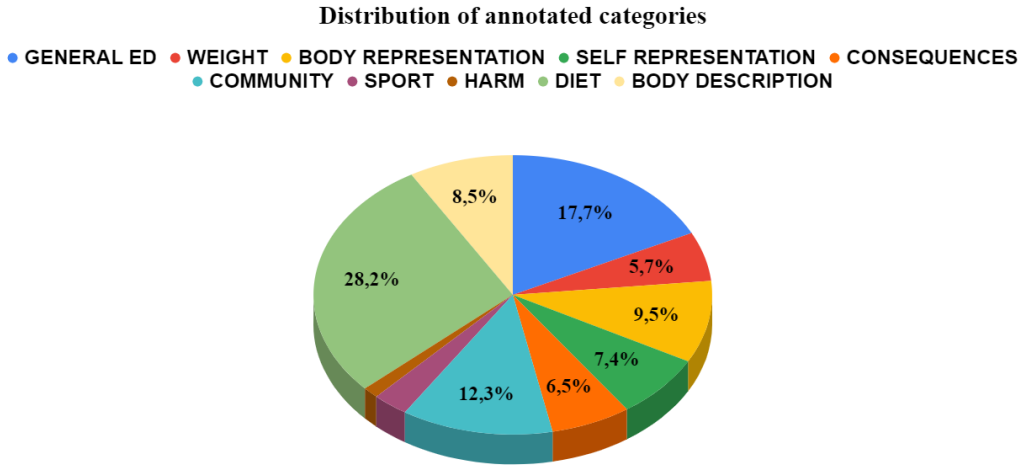

Figure 4 Annotations distribution.

end of the annotation task, we measured Cohen's Kappa, and it was 0.89 between the two annotators. We believe this is acceptable given the difficulty of the task. A set of rules and examples has been given to the annotators:

1. For each sentence they could choose up to two categories among those available;
2. When they found two categories, they could choose a main one and a secondary one, if necessary. Example: "*I feel so clean right now no food in a week #ProAna!*" has been annotated as primary DIET ("*no food in a week*") because the tweet is about not having food in a week and with the consequences of feeling clean, so the secondary will be SELF REPRESENTATION ("*I feel so clean*");
3. Sometimes Twitter users' use hashtags to index the content and have more chances to have their profile found. For this reason annotators distinguished hashtags between (a) unit of content and replaceable with the same word (example: "*I am happy to be in my #proana club!*"), and (b) – usually – a sequence at the end of the sentence used only to index the content and not relevant for the topic expressed (example: *"ugh! #proana #edtwt #ed #anabuddy #weightloss*").
4. Whenever it was not possible to label specific category and if the sentence was still related to ED they could use the GENERAL ED category.

A cross-reading has been done to improve the reliability of the dataset annotatation.

5.2 Annotation results and discussion

Table 6 Annotated sentences.

No. of sentences	**3,064**
No. of words	20,838
No. of types	5,528

By analyzing the annotation results (see Figure 4) we noticed three major categories: DIET, GENERAL ED, and COMMUNITY. We think that these results reflect the major characteristics of this community: their worries about what they eat (DIET), their need to have a group to whom to talk and share (COMMUNITY). Finally GENERAL ED regroups everything that may be shared online and not necessarily being confined in a specific category. This highlights also the wealth of arguments of this type of user.

We noticed that the most difficult sentences to annotate have been the ones that would be ambiguous without any further context, and the ones showing more than two categories at the same time, such as:

- "being *hungry asshat* until you get tired of it... *Die* FFS!": this sentence could be annotated as CONSEQUENCES (*being hungry*) SELF REPRESENTATION (*asshat*) HARM (*Die*);
- "*feeling fat*" versus "*be fat*": here annotators agreed to annotate the first as SELF REPRESENTATION and the second as BODY DESCRIPTION.

6 Word Mover's Distance to annotate similar sentences

In order to evaluate the use of Word Mover's Distance (WMD) for annotation, we used the same annotated dataset.

Word Mover's Distance is an adaption of Earth Mover's Distance (EMD) [36] which uses word embeddings to determine the similarity between two or more series of words (e.g., sentences). WMD uses the locations and words relative frequency weights of word embeddings to find the nearest neighbor for each word. Specifically, WMD uses the product of two numbers: the cosine distance between two words in the n-dimensional embedding space, and a weighting term that indicates how much one word in one document must travel to another word in the other document. In this way, it can minimize the cost of moving all words from a document to the positions of all the words in another document. The documents that share many semantically-similar words will have smaller distances than the documents with dissimilar words. In Figure 5, we show four sentences, originally from [24]:

1. D0: The President greets the press in Chicago
2. D1: Obama speaks to the media in Illinois
3. D2: The band gave a concert in Japan
4. D3: Obama speaks in Illinois

The relative cost of moving all the words in D2 to the locations of the words in D0 is greater than moving the words in documents D1 and D3. Formally:

$$WMD_{ij} = min_{T \geq 0} \sum_{i,j=1}^{2} T_{ij} c(i,j)$$

$$\sum_{j=1}^{2} T_{ij} = d_i, \forall i \in \{1, ..., n\}$$

$$\sum_{i=1}^{2} T_{ij} = d'_j, \forall j \in \{1, ..., n\}$$

with c(i,j) representing the euclidean distance $||\mathbf{x_i} - \mathbf{x_j}||_2$ between the two words in the embedding space. The travel cost between two words translates in the distance between texts. Let $\mathbf{d}$ and $\mathbf{d'}$ be the documents with each word i in $\mathbf{d}$ to be transformed into any words inside the document $\mathbf{d'}$. $\mathbf{T}$ is the sparse matrix where $\mathbf{d'}$ represents how much of i in $\mathbf{d}$ travels to word j in $\mathbf{d'}$. We expect that the moving from word i must equal to $\mathbf{d_i}$ in order to allow the transformation of $\mathbf{d}$ into $\mathbf{d'}$. The same is applicable for the word j that should match $\mathbf{d'_j}$.

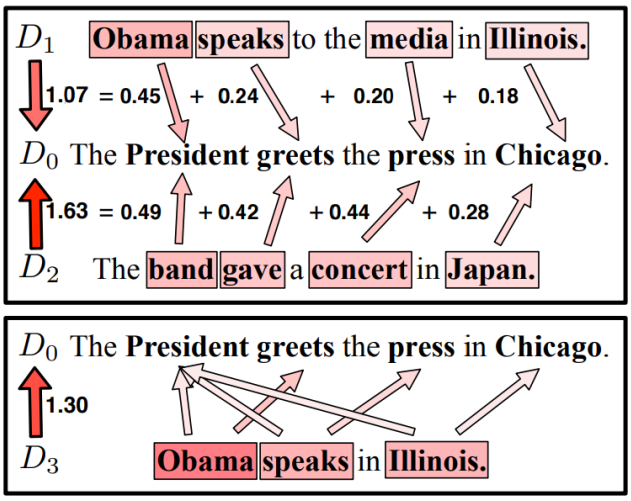

Figure 5 Illustration of Word Mover's Distance from Kusner et al. [24].

Our goal was to evaluate whether WMD could be used to improve the annotation process by verifying that the most similar sentences retrieved by WMD were of the right class. First of all, we trained Word2vec [30, 29] using the Gensim package [10] on the whole annotated dataset with vector size equal to 100. Then, we isolated 30% of the sentences for each class, and we run WMD against them. Finally, we retrieved the most similar sentences for each sentence of that 30% with a threshold of similarity of 0.98 and above, excluding the sentences used to compute the similarity. We evaluated this method by comparing the most similar sentences per class with their real class labels. Our results are shown on table 7. They show that WMD is a good model compared to CorEx to annotate new sentences when similarity is the searched parameter.

Table 7 F1 score for the sentences retrieved using Word Mover Distance.

	Weight	Body Representation	Self Representation	Harm	Consequences	Community	Sport	Diet	Body Description	General
F1-Score CorEx	0.0178	0.1291	0.0159	0.0275	0.0315	0.0676	0.0334	0.0564	0.1326	**0.6412**
F1-Score WMD	**0.7686**	**0.7909**	**0.7544**	**0.51**	**0.9721**	**0.8319**	**0.7404**	**0.7756**	**0.6114**	0.5955

7 Discussion and limitations

We acknowledge that this study is limited on several aspects: (a) people are self declaring to have an eating disorder, (b) within an online community, (c) expressing themselves according to the standard in use on Twitter and (d) we do not have any knowledge about their real life. However, we believe that they are representative of a part of people suffering from an eating disorder. A way to obtain more concrete results could be a joint study with clinical researchers in order to verify the validity of our study in a context outside the Internet and to improve it for other contexts. The annotation phase showed some limitations on the long run:

[10] https://radimrehurek.com/gensim/models/word2vec.html

- many human annotations fell under the GENERAL ED category found by CorEx: it could be possible to distinguish more topics that are a minority compared to others, but represent a big class altogether.
- Annotations on this domain, by human or by an algorithm, are complex: even though we decided only to use one label per sentence, we understand that there are some limitations, such as the co-presence of more topics in less than 20 words.

In the future, we would like to use both sentence similarities and a classification algorithm based on shallow parsing to capture them more accurately. We think that this could be improved by implementing syntactic rules (for example to capture implicit and explicit comparisons) and specific weights for words that are likely to be more in a category than in another. Take, for example, "burn"+"calorie" – we know that the word "calorie" is present in DIET, but the bigram "burn calorie" is likely to be more used in SPORT.

8 Conclusions

The main goals of this study were:

1. the creation of new resources, such as a pro-ED Twitter dataset and an annotated dataset both available under request – due to NDA, to facilitate and increase ED related studies on social media;
2. the exploration and sharing of alternative ways for the annotation experience, and the discovery of new keywords and textual items related to the studied issue, such as emojis, foreign language linguistic patterns, and uncommon use of words by employing two models: CorEx and WMD.

We believe that the work described in this paper can also be used in other online contexts where people's lives are in danger: suicide prevention, detection of depression signs, detection of harassment signs.

This work can be extended by using a classification framework to filter out dangerous expressions (encrypted or not), clustering them by topics, detecting keywords and increasing the number of keywords and topics. This will allow the early detection of possible online ED trends, such as the ABC diet and the Apple diet, or other dangerous online trends that were seen in the past (e.g., "*Blue Whale challenge*").

References

1 Alina Arseniev-Koehler, Hedwig Lee, Tyler McCormick, and Megan A Moreno. # proana: Pro-eating disorder socialization on Twitter. *Journal of Adolescent Health*, 58(6):659–664, 2016.

2 Carolina Figueras Bates. "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, 2015.

3 David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

4 Leah Boepple and J Kevin Thompson. A content analytic comparison of fitspiration and thinspiration websites. *International Journal of Eating Disorders*, 49(1):98–101, 2016.

5 Dina LG Borzekowski, Summer Schenk, Jenny L Wilson, and Rebecka Peebles. e-ana and e-mia: A content analysis of pro–eating disorder web sites. *American journal of public health*, 100(8):1526–1534, 2010.

6 Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Costello, Nina Kaiser, Elizabeth S Cahn, Ellen E Fitzsimmons-Craft, and Denise E Wilfley. "I just want to be skinny.": A content analysis of tweets expressing eating disorder symptoms. *PloS one*, 14(1):e0207506, 2019.

7 Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3213–3226, 2017.

8 Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1171–1184, 2016.

9 Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1201–1213, 2016.

10 Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 288–296. Curran Associates, Inc., 2009. URL: `https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf`.

11 Yohan Chon, Yungeun Kim, Hyojeong Shin, and Hojung Cha. Topic modeling-based semantic annotation of place using personal behavior and environmental features. *Transportation*, 23:110, 2009.

12 Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 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, 2015.

13 Kristen M Culbert, Sarah E Racine, and Kelly L Klump. Research review: What we have learned about the causes of eating disorders – a synthesis of sociocultural, psychological, and biological research. *Journal of Child Psychology and Psychiatry*, 56(11):1141–1164, 2015.

14 Munmun De Choudhury. Anorexia on tumblr: A characterization study. In *Proceedings of the 5th international conference on digital health 2015*, pages 43–50, 2015.

15 Elizabeth W Diemer, Julia D Grant, Melissa A Munn-Chernoff, David A Patterson, and Alexis E Duncan. Gender identity, sexual orientation, and eating-related pathology in a national sample of college students. *Journal of Adolescent Health*, 57(2):144–149, 2015.

16 Danielle A Gagne, Ann Von Holle, Kimberly A Brownley, Cristin D Runfola, Sara Hofmeier, Kateland E Branch, and Cynthia M Bulik. Eating disorder symptoms and weight and shape concerns in a large web-based convenience sample of women ages 50 and above: Results of the gender and body image (gabi) study. *International Journal of Eating Disorders*, 45(7):832–844, 2012.

17 Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, 2017.

18 Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542, 2017. `doi:10.1162/tacl_a_00078`.

19 Jannath Ghaznavi and Laramie D Taylor. Bones, body parts, and sex appeal: An analysis of# thinspiration images on popular social media. *Body image*, 14:54–61, 2015.

20 David Giles. Constructing identities in cyberspace: The case of eating disorders. *British journal of social psychology*, 45(3):463–477, 2006.

21 Grace J Johnson and Paul J Ambrose. Neo-tribes: The power and potential of online communities in health care. *Communications of the ACM*, 49(1):107–113, 2006.

22 Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint*, 2016. `arXiv:1612.03651`.

23 Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint*, 2016. arXiv:1607.01759.

24 Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.

25 David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk: early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 343–361. Springer, 2018.

26 David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer, 2019.

27 David E Losada, Fabio Crestani, and Javier Parapar. erisk 2020: Self-harm and depression challenges. In *European Conference on Information Retrieval*, pages 557–563. Springer, 2020.

28 Luana Marques, Margarita Alegria, Anne E Becker, Chih-nan Chen, Angela Fang, Anne Chosak, and Juliana Belo Diniz. Comparative prevalence, correlates of impairment, and service utilization for eating disorders across us ethnic groups: Implications for reducing ethnic disparities in health care access for eating disorders. *International Journal of Eating Disorders*, 44(5):412–420, 2011.

29 Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. arXiv:1301.3781.

30 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

31 David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262–272, USA, 2011. Association for Computational Linguistics.

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

33 Jessica A Pater, Brooke Farrington, Alycia Brown, Lauren E Reining, Tammy Toscos, and Elizabeth D Mynatt. Exploring indicators of digital self-harm with eating disorder patients: A case study. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

34 Danielle C Ransom, Jennifer G La Guardia, Erik Z Woody, and Jennifer L Boyd. Interpersonal interactions on online forums addressing eating concerns. *International Journal of Eating Disorders*, 43(2):161–170, 2010.

35 Azadeh A Rikani, Zia Choudhry, Adnan M Choudhry, Huma Ikram, Muhammad W Asghar, Dilkash Kajal, Abdul Waheed, and Nusrat J Mobassarah. A critique of the literature on etiology of eating disorders. *Annals of neurosciences*, 20(4):157, 2013.

36 Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998. doi:10.1109/ICCV.1998.710701.

37 Yuanlong Shao, Yuan Zhou, Xiaofei He, Deng Cai, and Hujun Bao. Semi-supervised topic modeling for image annotation. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, page 521–524, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1631272.1631346.

38 Shaina J Sowles, Monique McLeary, Allison Optican, Elizabeth Cahn, Melissa J Krauss, Ellen E Fitzsimmons-Craft, Denise E Wilfley, and Patricia A Cavazos-Rehg. A content analysis of an online pro-eating disorder community on reddit. *Body image*, 24:137–144, 2018.

39 Shabbir Syed-Abdul, Luis Fernandez-Luque, Wen-Shan Jian, Yu-Chuan Li, Steven Crain, Min-Huei Hsu, Yao-Chin Wang, Dorjsuren Khandregzen, Enkhzaya Chuluunbaatar, Phung Anh Nguyen, et al. Misleading health-related information promoted through video-based social media: anorexia on youtube. *Journal of medical Internet research*, 15(2):e30, 2013.

40 Catherine Victoria Talbot, Jeffrey Gavin, Tommy Van Steen, and Yvette Morey. A content analysis of thinspiration, fitspiration, and bonespiration imagery on social media. *Journal of eating disorders*, 5(1):1–8, 2017.

41 Marika Tiggemann, Owen Churches, Lewis Mitchell, and Zoe Brown. Tweeting weight loss: A comparison of# thinspiration and# fitspiration communities on Twitter. *Body Image*, 25:133–138, 2018.

42 Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF (Working Notes)*, 2018.

43 Suppawong Tuarob, Line C Pouchard, Prasenjit Mitra, and C Lee Giles. A generalized topic modeling approach for automatic document annotation. *International Journal on Digital Libraries*, 16(2):111–128, 2015.

44 Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, pages 577–585, 2014.

45 Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International conference on web search and data mining*, pages 91–100, 2017.

46 Elad Yom-Tov, Luis Fernandez-Luque, Ingmar Weber, and Steven P Crain. Pro-anorexia and pro-recovery photo sharing: a tale of two warring tribes. *Journal of medical Internet research*, 14(6):e151, 2012.

47 Wei Zhang, Yan-Chuan Sim, Jian Su, and Chew-Lim Tan. Entity linking with effective acronym expansion, instance selection and topic modeling. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

48 Sicheng Zhou, Yunpeng Zhao, Rubina Rizvi, Jiang Bian, Ann F Haynos, and Rui Zhang. Analysis of Twitter to identify topics related to eating disorder symptoms. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–4. IEEE, 2019.