Towards Using Word Embedding Vector Space for Better Cohort Analysis

Towards Using Word Embedding Vector Space for Better Cohort Analysis

Mohamed Bahgat[1], Steve Wilson[1], Walid Magdy[1,2]
[1]School of Informatics, The University of Edinburgh, Edinburgh, UK
[2]The Alan Turing Institute, London, UK
m.bahgat@sms.ed.ac.uk, steven.wilson@ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

On websites like Reddit, users join communities where they discuss specific topics which cluster them into possible cohorts. The authors within these cohorts have the opportunity to post more openly under the blanket of anonymity, and such openness provides a more accurate signal on the real issues individuals are facing. Some communities contain discussions about mental health struggles such as depression and suicidal ideation. To better understand and analyse these individuals, we propose to exploit properties of word embeddings that group related concepts close to each other in the embeddings space. For the posts from each topically situated sub-community, we build a word embeddings model and use handcrafted lexicons to identify emotions, values and psycholinguistically relevant concepts. We then extract insights into ways users perceive these concepts by measuring distances between them and references made by users either to themselves, others or other things around them. We show how our proposed approach can extract meaningful signals that go beyond the kinds of analyses performed at the individual word level.

Introduction

Social media is a rich source of textual content reflecting how people perceive different aspects of life. Authors tend to cluster around groups that discuss specific topics of interest to them. For example, users might follow and contribute to discussions surrounding specific hashtags in Twitter, participate in Groups in Facebook, or contribute to subreddits in Reddit. In case of Reddit, users have the opportunity to keep their anonymity, and posting anonymously encourages individuals to be more open about themselves. This can lead to more accurate signals of how individuals truly think about and perceive themselves and the world around them.

Given the rich textual resources available in the form of social media data, there has been ongoing research to apply different machine learning and natural language processing tools to help analyse (De Choudhury, Counts, and Horvitz 2013) and classify users discussing mental health related topics (De Choudhury et al. 2013), including cases where individuals are potentially in danger (Zirikly et al. 2019).

One type of widely used resource in these kinds of studies is word embeddings, (Mikolov et al. 2013) which can be trained on very large corpora, and then shared with other researchers and subsequently applied to many other tasks. Word embeddings are based on the encode-decode model where a model is trained using unsupervised corpus to either predict a word from context or predict context from word. These embeddings are also capable of encoding semantic information about words. For example, semantically similar words, such as colours or cities, are grouped together in the embeddings space. Embeddings can also encode transitive relations such as what a capital is to a country.

In this work, we propose using Word Embeddings to help analyse relations between different categories of concepts, emotions and entities in textual content extracted from social media. We apply this to the domain of mental health and use Reddit as our content source. We use lexicons to identify the different categories and extract clusters of embeddings that represent those. We assume that each subreddit represents a single cohort and we use the extracted clusters and the distance between them to identify cohort perceptions towards different aspects.

We make several contributions in this work. First, we present a novel approach in identifying user perceptions by extracting relations between concept clusters in the word embeddings space. We further refine by ranking categories with respect to distances relative to each other for each cohort. We then compare ranks with ones from a neutral, “control”, cohort to extract relations that are significant to other cohorts of interest.

Analysis and Classification of Mental States

There are many language clues that would allow us to understand and identify mental state and health issues for individuals. These clues can be extracted from either written or spoken communication.

One of the significant language clues is how we use self-reference. For example, prior work has shown that depressed and formerly depressed students used more first person singular pronouns (Rude, Gortner, and Pennebaker 2004). Compared to a general sample of poets, poets who died by suicide used more of these “I” words and fewer inclusive “we” words. However, both groups used high rates of negative emotional content, suggesting that this negative emotional signal alone is not enough to characterize suicidal tendencies (Stirman and Pennebaker 2001). Following the trend of the aforementioned studies, other work has shown that, in autobiographical memories, people with depression used more “I” words, more words related to the present tense and more words overall when recalling negative memories compared to when recalling positive memories (Himmelstein et al. 2018).

Individuals struggling with mental health can turn to social media to vent their worries. They can seek support from other users who might have gone through similar situations or interact with individuals who suffered related illnesses. Thus, social media has provided a generous resource for applying research of mental health. In fact, web data has been used to study suicide, depression, and anxiety since 2002 (Fekete 2002). Researchers using social media data have applied the same kinds of analyses that had been performed earlier in offline settings. For example, one study followed users that started on mental health subreddits, but eventually moved to post on */r/SuicideWatch*. The language of these users was marked by a higher number of verbs and adverbs, fewer nouns and entities, decreased readability scores, more “I” words, fewer “we”, “they”, “she/he” words, and longer, but fewer, posts overall (De Choudhury et al. 2016).

Handcrafted lexicons are popular in mental health applications. For example, the usage of psycholinguistically relevant categories from the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) is significantly different between users with versus without mental health disorders diagnoses, based on an analysis of Twitter data (Coppersmith, Dredze, and Harman 2014). The approach toward extracting information and classifying users has to be consistent to the application and avoid looking for the wrong signals that might not help. Neither sentiment nor emotional content was helpful in detecting suicide related communication on Twitter (Burnap, Colombo, and Scourfield 2015).

Analysis of mental health on social media has not been limited to language processing. A users’ individual interaction with their social network can also be used to classify their mental states. Based on interviews, it was found that non-depressed individuals actually perceived Twitter differently: as an information consuming and sharing tool, while depressed individuals perceived it as a tool for social awareness and emotional interaction (Park, McDonald, and Cha 2013). Further, increased social isolation was found to be a good predictor of postpartum depression using Facebook data (De Choudhury et al. 2014).

Hypotheses

Based on the studies mentioned previously, we outline the following hypotheses which we will test using our analysis method: *H1*: In text related to suicidal ideation, we expect to find a smaller distance between first person singular pronouns and death/negative emotion related words. *H2*: In text related to depression, we expect to find greater distances between first person singular pronouns and other pronouns.

Lexicons

To aid our analysis, we use handcrafted and crowd-created resources to extract the different signals and model psychological phenomena in textual data. Two lexicons are used to obtain wider list of categories for our investigation. These lexicons are Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. 2015) and a lexicon for measuring content related to personal values.

Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count is a tool that count number or percentage of word occurrences that belong to word categories. These categories where picked and validated by experts in the field and proved its usefulness in a wide range of applications (De Choudhury et al. 2013) (Van Der Lee et al. 2019). The lexicon contains word groups for self reference, others references, negative and positive emotions and different concepts such as death, health, risk. Categories are then grouped; for example, sadness, anger and anxiety are grouped into negative emotions category.

Values

The Values Lexicon is another hierarchical Lexicon, but the categories are all related to personal values that people hold (Wilson, Shen, and Mihalcea 2018). This lexicon was constructing using a mix of automated and human-in-the-loop methods to extract relevant words and arrange them into a hierarchical structure. It was later shown to be useful in studies of cross-cultural differences (Shen, Wilson, and Mihalcea 2019) and the analysis of user profiles on Twitter in a study of human activities (Wilson and Mihalcea 2019).

Data

We extract our data from the Social Media website Reddit, which is comprised of many forums, called Subreddits, dedicated to the discussion of specific topics. Subreddits can also be either public or private. Users can post their submissions and comment on their or other individuals’ submissions. Users may maintain anonymity as long as they do not expose any personal information on the website.

Following previous work that used Reddit data for mental health studies (De Choudhury and De 2014; Li, Mihalcea, and Wilson 2018), we use submissions from this set of mental health subreddits: */r/SuicideWatch*, */r/depression*, */r/mentalhealth*, */r/BPD*, */r/ptsd*, */r/bipolar2*, */r/rapecouncling*, */r/StopSelfHarm* and */r/socialanexity*. We have used only submissions from these Subreddits without their comments, as submissions would be more relevant to the authors themselves, while comments can be reaction to other authors submissions.

As a control group, we have chosen */r/IAmA*. This Subreddit contains wide range of topics that include mental health discussions as well as other general topics. With such generality, */r/IAmA* can be viewed as a neutral subreddit that contains a mix of different emotional as well as factual content compared to other subreddits such as */r/worldnews* which is biased towards factual content and has a more consistent way of posting in short sentences.

Subreddit	Submissions	Words	Vocab.
/r/depression	766,971	125,883,951	66,661
/r/SuicideWatch	296,647	49,142,594	44,552
/r/mentalhealth	106,750	16,177,175	31,853
/r/BPD	89,471	13,686,784	26,886
/r/socialanexity	74,802	9,674,364	22,041
/r/ptsd	21,545	4,194,729	17,042
/r/rapecouncling	14,907	4,401,250	14,488
/r/bipolar2	10,519	1,424,287	9,436
/r/StopSelfHarm	7,965	1,059,160	7,011
/r/IAmA	402,415	22,208,754	60,455

Table 1: Subreddit submission counts, word counts and vocabulary sizes.

Table 1 shows content statistics which was acquired from Pushshift (Baumgartner et al. 2020) and covered the period from January 2011 to August 2019.

Analysis using Cluster Relative Rank

To identify which concepts are relevant within a cohort, we estimate the distance between clusters that represent each concept. A small distance between two concepts signifies that these concepts are more semantically related from the point of view of authors in that cohort. For example, if the self-referential *LIWC:i*[1], cluster is close to the cluster representing the concept of *LIWC:death*, then it would suggest that authors within that cohort perceive themselves as somehow being associated with death. If the concept of *Values:Home* is close to *Values:Wealth*, then authors within that cohort may perceive that an established home is related to wealth while in another cohort where concepts of *Values:Home* and *Values:Family* are related together might indicate that authors are more attached to family at home.

Using word embeddings allows semantic relations to be encoded (Mikolov et al. 2013). Simply counting words from LIWC lexicon can extract multiple concepts that exist in the cohort corpus, but would not reveal which ones are related to which. This can be solved by co-occurrence statistics, yet this approach will lack the deeper semantics provided by word embeddings that makes use of relationships between words that aren't included in the lexicon but exist in the corpus.

Approach

For each subreddit, we build a word embeddings model using only submissions for that given subreddit. The corresponding vectors from each category of words from LIWC and Values are identified. Each cluster is then represented by the centroid of the vector values. We follow (Mikolov et al. 2013) in using cosine distances to measure distances the embeddings space. The cosine distance between each centroid is calculated between each pair of categories. Relative to each category, we rank other categories based on how close they are. Categories ranked at the top would be the most relevant or highly associated for the corresponding concept from the perspective of individuals from the cohort that authored the corpus of the subreddit. We have also noticed that comparing ranks to a control subreddit that is neutral (in our case it is */r/IAmA*) gives clearer results as this comparison eliminates categories that are commonly used on social media but generally irrelevant to our study, such as *LIWC:netspeak* and *LIWC:informal*. Thus our outcome is sorted with respect to change of rank of a category relevant to another comparing its rank within the current subreddit of interest to the rank in the control subreddit.

More formally, for each subreddit r, we train a distinct set of word embeddings E_r, where E_r^w the embedding of the word w in E_r. In each embedding space, we compute the average embedding (centroid) for a lexicon category, $X_r(C) = \sum_{c \in C} E_r^c / |C|$. Let $d(r, C_1, C_2) = \cos\_\mathrm{dist}(X_r(C_1), X_r(C_2))$, represent cosine distance between the average embeddings of categories C_1 and C_2 in the embedding space of E_r. This value illustrates, for this subreddit, how semantically similar the group of words in C_1 is to the group of words in C_2. We then compare these values across subreddits in order to determine which categories have moved closer or farther away from each other, representing this quantity as $\Delta(r_1, r_2, C_1, C_2) = d(r_1, C_1, C_2) - d(r_2, C_1, C_2)$. Higher values indicate that, compared to r_1, distance between C_1 and C_2 is lower in r_2.

Subreddits with Smaller Corpus Size

For some subreddits, the corpus size that we have collected was considerably small. To verify the effect of size, we have built four different models from */r/StopSelfHarm* (one of the smaller subreddits), computed category centroids, and then calculated distances between each pair. The variance in results between the four models was considerably small; in the order of $1e^{-5}$, suggesting that these inter-centroid distances are relatively stable for the size of corpus that we are using.

Results

We have used fasttext (Joulin et al. 2016) to generate skip-gram word embeddings for each subreddit. The vector embeddings size was set to 100.

The distances between centroids representing categories vary for each cohort represented by a mental related subreddit. For instance, for people with suicidal ideation posting on */r/SuicideWatch*, *LIWC:death* category appeared closer to self reference category *LIWC:i* compared to other subreddits as shown in Figure 1. For example, users mentioning graves with relation to them personally such as *"It is time the grave swallow me up. I have lived enough."* or *"I feel like if I move I'll walk out to my dog's grave and lie down with him forever."* On the other hand, for bipolar disorder */r/BPD*, both *LIWC:body* and *LIWC:feel* appeared close to the self reference centroid as shown in Figure 2; for example, *"I feel really guilty for feeling this way, but I can't help it."*. While there have been several mentions of suicidal tendencies; such as *I've been extremely suicidal since then and have gone through four psychiatrist"*, on the other hand, the

[1] We use the notation *Lexicon-name:category-name* to identify categories across lexicons.

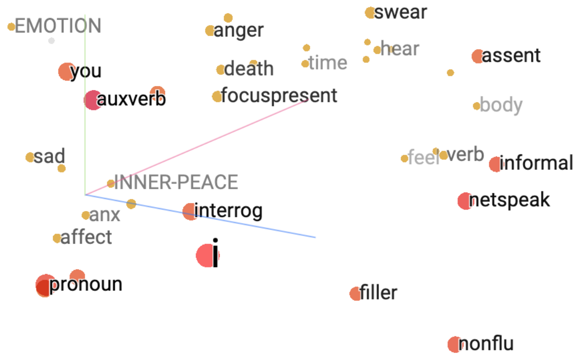

Figure 1: */r/SuicideWatch* centroids PCA projection into 3D space. Darker colours signifies shorter distance to *LIWC:i*. In original space, cosine distance between *LIWC:i* and *LIWC:death* is 0.309, while *LIWC:i* and *LIWC:negemo* has 0.327

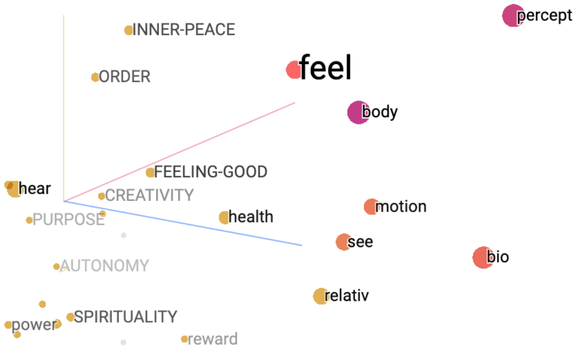

Figure 2: */r/BPD* centroids PCA projection. Darker colours signifies shorted distance to *LIWC:feel*. In original space, cosine distance between *LIWC:feel* and its closest neighbor *LIWC:body* is 0.053, while *LIWC:death* is the 30th closest neighbour with distance 0.252.

percentage of people expressing these thoughts are less compared to the ones in */r/SuicideWatch*.

The next step is to compute category rank difference between mental related subreddits and our control subreddit. We cross-reference our results against the hypotheses mentioned earlier. For the first case, where death becomes more relevant to self-reference for suicidal ideation, *LIWC:death* jumped 30 places (fourth biggest positive change) towards being more relevant for */r/SuicideWatch* users compared to */r/IAmA* users. Table 2 shows a sample of rank changes.

To investigate further, we have also compared distances with respect to plural first person reference; *LIWC:we*. While *LIWC:death* jumped 25 places when comparing relevance ranks between */r/SuicideWatch* and */r/IAmA* but that was the 23rd biggest jump for *LIWC:we* category. Also, the cosine distance between *LIWC:i* and *LIWC:death* centroids were smaller when compared to the distance *LIWC:we* and *LIWC:we*; 1.055 versus 1.716 respectively.

Category	RG	Dist.	Category	RG	Dist.
Values:Animals	41	0.44	*LIWC:negate*	27	0.34
LIWC:focusfuture	36	0.38	*Values:Friends*	24	0.43
LIWC:number	34	0.46	*Values:Cog.*	24	0.44
LIWC:death	30	0.31	*LIWC:they*	23	0.35
LIWC:nonflu	27	0.22	*LIWC:relig*	22	0.42

Table 2: Top 10 rank gaining (RG) categories with respect to self reference category *LIWC:i* when comparing */r/SuicideWatch* with *IAmA*. The cosine distance within the same embeddings space for */r/SuicideWatch* is also shown.

Category 1	Category 2	Distance	Rank Gain
LIWC:i	*LIWC:sad*	0.338	7
LIWC:we	*LIWC:sad*	0.524	-2
LIWC:i	*LIWC:negemo*	0.358	-1
LIWC:we	*LIWC:negemo*	0.563	-5

Table 3: Cosine distances between *LIWC:i* versus *LIWC:we* for categories relevant to depression in */r/depression* with closest neighbour rank gains compared to */r/IAmA*

In our second hypothesis, individuals become more isolated due to depression. For individuals within that cohort, self-reference is being associated more with independence; that is *Values:Autonomy* moved up seven places while other associations that would indicate social and open behavior were less relevant to self-reference compared to our control set. Some of these categories were *LIWC:you* down one place, *Values:Self Confidence* four places, *Values:Relationships* five places, *Values:Optimism* five places, *Values:Marriage* five places, *LIWC:friend* eleven places, and *LIWC:family* 20 places.

Again, we compare cosine distances of first person singular; *LIWC:i*; versus first person plural; *LIWC:we*, with respect to two depression related categories which are sadness; *LIWC:sad* and negative emotions *LIWC:negemo* for */r/depression* subreddit. Table 3 shows that first person singular is closer to both depression related categories within the same space representing */r/depression* compared to first person plural. Also, the table shows when comparing to */r/IAmA*; our control subreddit, rank gains were higher again for both categories for first person singular.

One interesting phenomenon was that *Values:Animal* category appeared at a high relative rank when associated with *LIWC:i* self-reference for most of mental subreddits though our control set contained considerable number of mentions for that category. Looking at submissions text, mentions of animals were either to refer to humans; such as in *"I feel worthless, useless, and like a horrible, vile creature that the whole world must hate."* or *"She was the most beautiful creature I have ever seen."*, or to mention that they own pets; such as in *"I have a pet guinea pig, she's the only thing that even remotely brings me joy."* or *"I was so sad to leave my home and pets"*. The later case would resonate with some studies implying that interaction with pets and animals is helpful for mental disorder (Souter and Miller 2007).

Conclusions

In this work, we presented a novel method of analysing social content with the focus on mental related content. Our work exploited the semantics encoded within word embeddings to identify relations between different concepts and emotions from the perspective of authors within a specified cohort. While general relations surfaced within our results, comparing with a control corpus help suppress some of these. We believe this method has potential moving forward to improve the way analysis is being done. It can also help picking the right categories to use in classification or filter content to improve classification accuracy.

Acknowledgements

This work was partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435*.

Burnap, P.; Colombo, W.; and Scourfield, J. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*, 75–84. ACM.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 51–60.

De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

De Choudhury, M.; Counts, S.; Horvitz, E. J.; and Hoff, A. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 626–638. ACM.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110. ACM.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 47–56. ACM.

Fekete, S. 2002. The internet-a new source of data on suicide, depression and anxiety: A preliminary study. *Archives of Suicide Research* 6(4):351–361.

Himmelstein, P.; Barb, S.; Finlayson, M. A.; and Young, K. D. 2018. Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PloS one* 13(11):e0207814.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Li, Y.; Mihalcea, R.; and Wilson, S. R. 2018. Text-based detection and understanding of changes in mental health. In *International Conference on Social Informatics*, 176–188. Springer.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Park, M.; McDonald, D. W.; and Cha, M. 2013. Perception differences between the depressed and non-depressed users in twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of liwc2015. Technical report.

Rude, S.; Gortner, E.-M.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18(8):1121–1133.

Shen, Y.; Wilson, S. R.; and Mihalcea, R. 2019. Measuring personal values in cross-cultural user-generated content. In *International Conference on Social Informatics*, 143–156. Springer.

Souter, M. A., and Miller, M. D. 2007. Do animal-assisted activities effectively treat depression? a meta-analysis. *Anthrozoös* 20(2):167–180.

Stirman, S. W., and Pennebaker, J. W. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine* 63(4):517–522.

Van Der Lee, C.; Van Der Zanden, T.; Krahmer, E.; Mos, M.; and Schouten, A. 2019. Automatic identification of writers' intentions: Comparing different methods for predicting relationship goals in online dating profile texts. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 94–100.

Wilson, S., and Mihalcea, R. 2019. Predicting human activities from user-generated content. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2572–2582.

Wilson, S. R.; Shen, Y.; and Mihalcea, R. 2018. Building and validating hierarchical lexicons with a case study on personal values. In *International Conference on Social Informatics*, 455–470. Springer.

Zirikly, A.; Resnik, P.; Uzuner, Ö.; and Hollingshead, K. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.