RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses

Sean MacAvaney*, Bart Desmet†*, Arman Cohan*, Luca Soldaini*,
Andrew Yates‡*, Ayah Zirikly§, Nazli Goharian*

*IR Lab, Georgetown University, US
{firstname}@ir.cs.georgetown.edu

†LT3, Ghent University, BE
bart.desmet@ugent.be

‡Max Planck Institute for Informatics, DE
ayates@mpi-inf.mpg.de

§ National Institutes of Health, US
ayah.zirikly@nih.gov

Abstract

Self-reported diagnosis statements have been widely employed in studying language related to mental health in social media. However, existing research has largely ignored the temporality of mental health diagnoses. In this work, we introduce RSDD-Time: a new dataset of 598 manually annotated self-reported depression diagnosis posts from Reddit that include temporal information about the diagnosis. Annotations include whether a mental health condition is present and how recently the diagnosis happened. Furthermore, we include exact temporal spans that relate to the date of diagnosis. This information is valuable for various computational methods to examine mental health through social media because one's mental health state is not static. We also test several baseline classification and extraction approaches, which suggest that extracting temporal information from self-reported diagnosis statements is challenging.

1 Introduction

Researchers have long sought to identify early warning signs of mental health conditions to allow for more effective treatment (Feightner and Worrall, 1990). Recently, social media data has been utilized as a lens to study mental health (Coppersmith et al., 2017). Data from social media users who are identified as having various mental health conditions can be analyzed to study common language patterns that indicate the condition; language use could give subtle indications of a person's wellbeing, allowing the identification of at-risk users. Once identified, users could be provided with relevant resources and support.

While social media offers a huge amount of data, acquiring manually-labeled data relevant to mental health conditions is both expensive and not scalable. However, a large amount of labeled data is crucial for classification and large-scale analysis. To alleviate this problem, NLP researchers in mental health have used unsupervised heuristics to automatically label data based on self-reported diagnosis statements such as "I have been diagnosed with depression" (De Choudhury et al., 2013; Coppersmith et al., 2014a, 2015; Yates et al., 2017).

A binary status of a user's mental health conditions does not tell a complete story, however. People's mental condition changes over time (Wilkinson and Pickett, 2010), so the assumption that language characteristics found in a person's social media posts historically reflects their current state is invalid. For example, the social media language of an adult diagnosed with depression in early adolescence might no longer reflect any depression. Although the extraction of temporal information has been well-studied in the clinical domain (Lin et al., 2016; Bethard et al., 2017; Dligach et al., 2017), temporal information extraction has remained largely unexplored in the mental health domain. Given the specific language related to self-reported diagnoses posts and the volatility of mental conditions in time, the time of diagnosis provides critical signals on examining mental health through language.

To address this shortcoming of available datasets, we introduce RSDD-Time: a dataset of temporally annotated self-reported diagnosis statements, based on the Reddit Self-Reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017). RSDD-Time includes 598 diagnosis statements that are manually annotated to include pertinent temporal information. In particular, we identify if the conditions are current, meaning that the condition is apparently present according the the

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 168–173
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

self-reported diagnosis post. Next, we identify how recently a particular diagnosis has occurred. We refer to these as *condition state* and *diagnosis recency*, respectively. Furthermore, we identify the time expressions that relate to the diagnosis, if provided.

In summary, our contributions are: *(i)* We explain the necessity of temporal considerations when working with self-reported diagnoses. *(ii)* We release a dataset of annotations for 598 self-reported depression diagnoses. *(iii)* We provide and analyze baseline classification and extraction results.

Related work Public social media has become a lens through which mental health can be studied as it provides a public narration of user activities and behaviors (Conway and O'Connor, 2016). Understanding and identifying mental health conditions in social media (e.g., Twitter and Reddit) has been widely studied (De Choudhury et al., 2013; Coppersmith et al., 2014b; De Choudhury and De, 2014; Mitchell et al., 2015; Gkotsis et al., 2016; Yates et al., 2017). To obtain ground truth knowledge for mental health conditions, researchers have used crowdsourced surveys and heuristics such as self-disclosure of a diagnosis (De Choudhury et al., 2013; Tsugawa et al., 2015). The latter approach uses high-precision patterns such as "I was diagnosed with depression." Only statements claiming an actual diagnosis are considered because people sometimes use phrases such as "I am depressed" casually. In these works, individuals self-reporting a depression diagnoses are presumed to be depressed. Although the automated approaches have yielded far more users with depression than user surveys (tens of thousands, rather than hundreds), there is no indication of whether or not the diagnosis was recent, or if the conditions are still present. In this work, we address this by presenting manual annotations of nearly 600 self-reported diagnosis posts. This dataset is valuable because it allows researchers to train and test systems that automatically determine diagnosis recency and condition state information.

2 Data

For the study of temporal aspects of self-reported diagnoses, we develop an annotation scheme[1] and apply it to a set of 598 diagnosis posts randomly sampled from the Reddit Self-Reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017). In the annotation environment, the diagnosis match is presented with a context of 300 characters on either side. A window of 150 characters on either side was too narrow, and having the whole post as context made annotation too slow, and rarely provided additional information.

Annotation scheme Two kinds of text spans are annotated: diagnoses (e.g., "I was diagnosed") and time expressions that are relevant to the diagnosis (e.g., "two years ago"). On diagnosis spans, the following attributes are marked:

- **Diagnosis recency** determines when the diagnosis occurred (not the onset of the condition). Six categorical labels are used: very recently (up to 2 months ago), more than 2 months but up to 1 year ago, more than 1 year but up to 3 years ago, more than 3 years ago, *unspecified* (when there is no indication), and *unspecified but not recent* (when the context indicates that the diagnosis happened in the past, yet there is insufficient information to assign it to the first four labels).
- For **condition state**, the annotator assesses the context for indications of whether the diagnosed condition is still current or past. The latter includes cases where it is reported to be fully under control through medication. We use a five-point scale (*current, probably current, unknown, probably past* and *past*). This can be mapped to a three-point scale for coarse-grained prediction (i.e. moving *probable* categories to the center or the extremes).
- When a diagnosis is presented as uncertain or incorrect, we mark it as **diagnosis in doubt**. This can be because the diagnosis is put into question by the poster (e.g., "I was diagnosed with depression before they changed it to ADHD"), or it was later revised.
- Occasionally, incorrect diagnosis matches are found in RSDD. These are marked as **false positive**. This includes diagnoses for conditions other than depression or self-diagnosis that occur in block quotes from other posts. False positive posts are not included in the analyses below.

Time expressions indicating the time of diagnosis are marked similarly to the TIMEX3 specification (Pustejovsky et al., 2005), with the additional

[1] Available at `https://github.com/Georgetown-IR-Lab/RSDD-Time`

Span	Attribute	%	κ
diagnosis	false positive	0.97	0.43
	diagnosis in doubt	0.97	0.22
	condition state	0.52	0.41
	diagnosis recency	0.66	0.64
time	explicit	0.91	0.81
	inferable from age	0.93	0.82

Table 1: Inter-annotator agreement by average pairwise agreement (%) and weighted Cohen's kappa (κ).

support for ages, years in school, and references to other temporal anchors. Because of these additions, we also annotate prepositions pertaining to the temporal expression when present (e.g., 'at 14', 'in 2004'). Each span also has an indication of how their associated diagnosis can be assigned to one of the *diagnosis recency* labels. **Explicit** time expressions allow immediate assignment given the post date (e.g., yesterday, last August, in 2006). If the recency can be inferred assuming a poster's age at post time is known, it is **inferable from age** (e.g., at 17, in high school). A poster's age could be established using mentions by the author, or estimated with automatic age prediction.

Inter-annotator agreement After an initial annotation round with 4 annotators that allowed for the scheme and guidelines to be improved, the entire dataset was annotated by 6 total annotators with each post being at least double annotated; disagreements were resolved by a third annotator where necessary. We report pairwise inter-annotator agreement in Table 1. Cohen's kappa is linearly weighted for ordinal categories (*condition state* and *diagnosis recency*).

Agreement on false positives and doubtful diagnoses is low. For future analyses that focus on detecting potential misdiagnoses, further study would be required to improve agreement, but it is tangential to the focus on temporal analysis in this study.

Estimating the state of a condition is inherently ambiguous, but agreement is moderate at 0.41 weighted kappa. The five-point scale can be backed off to a three-point scale, e.g. by collapsing the three middle categories into *don't know*. Pairwise percent agreement then improves from 0.52 to 0.68. The recency of a diagnosis can be established with substantial agreement ($\kappa = 0.64$). Time expression attributes can be annotated with almost perfect agreement.

Attribute	Count
false positive	25 out of 598
diagnosis in doubt	16 out of remaining 573
condition state	current (254), prob. current (64), unknown (225), prob. past (29), past (26)
diagnosis recency	unspec. (232), unspec. but past (176), recent (27), >2m-1y (37), >1y-3y (29), >3y (97)
time expression	explicit (144), inferable from age (101), non-inferable (47), n/a (306)

Table 2: Attribute counts in the RSDD-Time dataset.

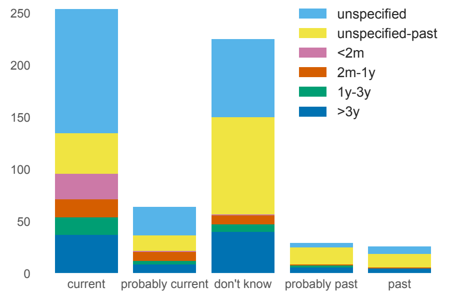

Figure 1: Incidence and interaction of *condition state* (columns) and *diagnosis recency* (colors).

Availability The annotation data and annotation guidelines are available at https://github.com/Georgetown-IR-Lab/RSDD-Time. The raw post text is available from the RSDD dataset via a data usage agreement (details available at http://ir.cs.georgetown.edu/resources/rsdd.html).

3 Corpus analysis

Counts for each attribute are presented in Table 2. Figure 1 shows the incidence and interaction between *condition state* and *diagnosis recency* in our dataset. About half the cases have a *condition state* that is current, but interestingly, there are also many cases (55) where the diagnosis relates (at least probably) to the past. There is also a large number of cases (225) where it is not clear from the post whether the condition is current or not. This further shows that many self-reported diagnosis statements may not be current, which could make a dataset noisy, depending on the objective. For *diagnosis recency*, we observe that the majority of diagnosis times are either unspecified or happened in the unspecified past. For 245 cases, however, the *diagnosis recency* can be inferred from the post, usually because there is an explicit time expression (59% of cases), or by inferencing from age (41%). Next, we investigate the interaction between *condition state* and *diagnosis recency*. We particularly observe that the majority of past conditions (rightmost two columns) are also associated with a *diagnosis recency* of more than 3 years ago or of an unspecified past. On the other hand, many current conditions (leftmost column) have an unspecified diagnosis time. This is expected because individuals who specifically indicate that their condition is not current also tend to specify when they have been first diagnosed, whereas individuals with current conditions may not mention their time of diagnosis.

4 Experiments

To gain a better understanding of the data and provide baselines for future work to automatically perform this annotation, we explore methods for attribute classification for *diagnosis recency* and *condition state*, and rule-based diagnosis time extraction. We split the data into a training dataset (399 posts) and a testing dataset (199 posts). We make this train/test split available for future work in the data release. For our experiments, we then disregard posts that are labeled as *false positive* (yielding 385 posts for training and 188 for testing), and we only consider text in the context window with which the annotator was presented.

4.1 Diagnosis recency and condition state classification

We train several models to classify *diagnosis recency* and *condition state*. In each we use basic bag-of-character-ngrams features. Character ngrams of length 2-5 (inclusive) are considered, and weighted using *tf-idf*. For labels, we use the combined classes described in Section 2. To account for class imbalance, samples are weighed by the inverse frequency of their category in the training set.

We compare three models: logistic regression, a linear-kernel Support Vector Machine (SVM), and Gradient-Boosted ensemble Trees (GBT) (Chen and Guestrin, 2016). The logistic regression and SVM models are ℓ_2 normalized, and the GBT models are trained with a maximum tree depth of 3 to avoid overfitting.

We present results in Table 3. The GBT method performs best for *diagnosis recency* classification, and logistic regression performs best for *condition state* classification. This difference could be due to differences in performance because of skew. The *condition state* data is more skewed, with *current* and *don't know* accounting for almost 80% of the labels.

	Diagnosis Recency			**Condition State**		
	P	**R**	**F1**	**P**	**R**	**F1**
Logistic Reg.	0.47	0.35	0.37	0.45	**0.45**	**0.44**
Linear SVM	0.23	0.23	0.21	**0.68**	0.40	0.40
GBT	**0.56**	**0.42**	**0.46**	0.35	0.38	0.36

Table 3: Macro-averaged classification results for *diagnosis recency* and *condition state* using *tf-idf* vectorized features for various baseline models.

4.2 Time expression classification

To automatically extract time expressions, we use the rule-based SUTime library (Chang and Manning, 2012). Because diagnoses often include an age or year in school rather than an absolute time, we added rules specifically to capture these time expressions. The rules were manually generated by examining the training data, and will be released alongside the annotations.

RSDD-Time temporal expression annotations are only concerned with time expressions that relate to the diagnosis, whereas SUTime extracts all temporal expressions in a given text. We use a simple heuristic to resolve this issue: simply choose the time expression closest to the post's diagnosis by character distance. In the case of a tie, the heuristic arbitrarily selects the leftmost expression. This heuristic will improve precision by eliminating many unnecessary temporal expressions, but has the potential to reduce precision by eliminating some correct expressions that are not the closest to the diagnosis.

Results for temporal extraction are given in Table 4. Notice that custom age rules greatly improve the recall of the system. The experiment also shows that the *closest* heuristic improves precision at the expense of recall (both with and without the age rules). Overall, the best results in terms of F1 score are achieved using both the *closest* heuristic and the age rules. A more sophisticated algorithm could be developed to increase the candidate expression set (to improve recall), and better predict which temporal expressions likely correspond to the diagnosis (to improve precision).

	P	R	F1
SUTime	0.17	0.59	0.26
+ age rules	0.20	**0.81**	0.32
+ closest heuristic	0.33	0.51	0.40
+ closest heuristic + age rules	**0.44**	0.69	**0.53**

Table 4: Results using SUTime, with additional rules for predicting age expressions and when limiting the candidate expression set using the *closest* heuristic.

5 Conclusion

In this paper, we explained the importance of temporal considerations when working with language related to mental health conditions. We introduced RSDD-Time, a novel dataset of manually annotated self-reported depression diagnosis posts from Reddit. Our dataset includes extensive temporal information about the diagnosis, including when the diagnosis occurred, whether the condition is still current, and exact temporal spans. Using RSDD-Time, we applied rule-based and machine learning methods to automatically extract these temporal cues and predict temporal aspects of a diagnosis. While encouraging, the experiments and dataset allow much room for further exploration.

References

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Angel X Chang and Christopher D Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *LREC*, volume 2012, pages 3735–3740.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, pages 393–396. IEEE.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain. Association for Computational Linguistics.

John W Feightner and Graham Worrall. 1990. Early detection of depression by primary care physicians. *CMAJ: Canadian Medical Association Journal*, 142(11):1215.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 63–73, San Diego, CA, USA. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2016. Improving temporal relation extraction with training instance augmentation. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 108–113, Berlin, Germany. Association for Computational Linguistics.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language timeml. *The language of time: A reader*, pages 545–557.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *CHI*.

Richard Wilkinson and Kate Pickett. 2010. *The spirit level: Why equality is better for everyone*. Penguin UK.

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