Improving Mental Health Classifier Generalization with Pre-Diagnosis Data

Yujian Liu, Laura Biester, Rada Mihalcea
Language and Mental Health

Linguistic patterns of depressed people have been studied using varied sources of data. Some linguistic patterns are more prevalent in depressed people’s speech and writing. Depressed people display self-focus (I-words) and use more negative emotion words (e.g., about anxiety and sadness).

Many researchers have used social media data to build depression classifiers. A BIG challenge: high quality training data.

Rude et al. (2004), Language use of depressed and depression-vulnerable college students
Rodriguez et al. (2010), Reading between the lines: the lay assessment of subclinical depression from written self-descriptions
Eichstaedt et al. (2018), Facebook language predicts depression in medical records
Language and Mental Health

• Linguistic patterns of depressed people have been studied using varied sources of data

Rude et al. (2004), Language use of depressed and depression-vulnerable college students
Rodriguez et al. (2010), Reading between the lines: the lay assessment of subclinical depression from written self-descriptions
Eichstaedt et al. (2018), Facebook language predicts depression in medical records
Linguistic patterns of depressed people have been studied using varied sources of data.

Some linguistic patterns are more prevalent in depressed people’s speech and writing:

- Depressed people display self-focus (I-words) and use more negative emotion words (e.g., about anxiety and sadness).

Many researchers have used social media data to build depression classifiers.

BIG challenge: high quality training data.

Rude et al. (2004), Language use of depressed and depression-vulnerable college students
Rodriguez et al. (2010), Reading between the lines: the lay assessment of subclinical depression from written self-descriptions
Eichstaedt et al. (2018), Facebook language predicts depression in medical records
Language and Mental Health

• Linguistic patterns of depressed people have been studied using varied sources of data

• Some linguistic patterns are more prevalent in depressed people’s speech and writing
  • Depressed people display self-focus (I-words) and use more negative emotion words (e.g., about anxiety and sadness)

• Many researchers have used social media data to build depression classifiers

Rude et al. (2004), Language use of depressed and depression-vulnerable college students
Rodriguez et al. (2010), Reading between the lines: the lay assessment of subclinical depression from written self-descriptions
Eichstaedt et al. (2018), Facebook language predicts depression in medical records
Language and Mental Health

• Linguistic patterns of depressed people have been studied using varied sources of data

• Some linguistic patterns are more prevalent in depressed people’s speech and writing
  • Depressed people display *self-focus* (I-words) and use more negative emotion words (e.g., about *anxiety* and *sadness*)

• Many researchers have used social media data to build depression classifiers
  • **BIG challenge:** high quality training data

Rude et al. (2004), Language use of depressed and depression-vulnerable college students
Rodriguez et al. (2010), Reading between the lines: the lay assessment of subclinical depression from written self-descriptions
Eichstaedt et al. (2018), Facebook language predicts depression in medical records
I need help. A few years ago, I was diagnosed with depression, which is common in my family. Anti-depressants helped for a while, but I am no longer able to use them... I have begun to [description of self-harm] again, and I hate doing it but can't stop. I hate my job, and I have nobody to support me, especially not my family. I simply don't know what to do. Thank you for reading this...

Coppersmith et al. (2014), Quantifying Mental Health Signals in Twitter
We refer to a statement such as “I have been diagnosed with depression” as a self-report.

I need help. A few years ago, I was diagnosed with depression, which is common in my family. Anti-depressants helped for a while, but I am no longer able to use them... I have begun to [description of self-harm] again, and I hate doing it but can't stop. I hate my job, and I have nobody to support me, especially not my family. I simply don't know what to do. Thank you for reading this...

Coppersmith et al. (2014), Quantifying Mental Health Signals in Twitter
I need help. A few years ago, I was diagnosed with depression, which is common in my family. Anti-depressants helped for a while, but I am no longer able to use them... I have begun to [description of self-harm] again, and I hate doing it but can't stop. I hate my job, and I have nobody to support me, especially not my family. I simply don't know what to do. Thank you for reading this...

Coppersmith et al. (2014), Quantifying Mental Health Signals in Twitter

- We refer to a statement such as “I have been diagnosed with depression” as a **self-report**
- Self-report patterns are commonly used to collect diagnosis labels for social media users
I need help. A few years ago, I was diagnosed with depression, which is common in my family. Anti-depressants helped for a while, but I am no longer able to use them... I have begun to [description of self-harm] again, and I hate doing it but can't stop. I hate my job, and I have nobody to support me, especially not my family. I simply don't know what to do. Thank you for reading this...

Coppersmith et al. (2014), Quantifying Mental Health Signals in Twitter

• We refer to a statement such as “I have been diagnosed with depression” as a self-report
• Self-report patterns are commonly used to collect diagnosis labels for social media users
• Their other posts are collected to train classifiers
Self-Reports and Generalization

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?

A person who has reported their mental health diagnosis on social media
Self-Reports and Generalization

A person who has reported their mental health diagnosis on social media

A person who has a diagnosed mental health condition

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Self-Reports and Generalization

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigan et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Self-Reports and Generalization

• Users who self-report aren't representative of the full population

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals

Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Self-Reports and Generalization

• Users who self-report aren't representative of the full population

• Models built using datasets based on self-report don’t generalize well

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Self-Reports and Generalization

• Users who self-report aren't representative of the full population

• Models built using datasets based on self-report don’t generalize well

A person who has *reported* their mental health diagnosis on social media

A person who has a *diagnosed* mental health condition

A person who has an *undiagnosed* mental health condition

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Self-Reports and Generalization

• Users who self-report aren't representative of the full population

• Models built using datasets based on self-report don’t generalize well

A person who has reported their mental health diagnosis on social media
A person who has a diagnosed mental health condition
A person who has an undiagnosed mental health condition

How can we improve generalization?

Ernala et al. (2019), Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals
Harrigian et al. (2020), Do Models of Mental Health Based on Social Media Data Generalize?
Reporting Timeline

symptom onset  seeking help  formal diagnosis

Time
Reporting Timeline

- Symptom onset
- Seeking help
- Formal diagnosis

Period in which a valid self-report could happen
Reporting Timeline

- **Symptom Onset**: Period in which people need to be connected to support.
- **Seeking Help**: Time period.
- **Formal Diagnosis**: Period in which a valid self-report could happen.

**Time**
Reporting Timeline

- **Period considered in training data based on self-report**
- **Period in which people need to be connected to support**
- **Period in which a valid self-report could happen**

- **Symptom onset**
- **Seeking help**
- **Formal diagnosis**

**time**
Reporting Timeline

- Symptom onset
- Seeking help
- Formal diagnosis
- Period in which people need to be connected to support
- Period in which a valid self-report could happen

Users who self-report were once in one of these stages.
Experiments
Experiments

**In-Domain:**

Does model performance drop when tested on pre-diagnosis data rather than data from all time periods?
Experiments

**In-Domain:**
Does model performance drop when tested on pre-diagnosis data rather than data from all time periods?

**Out-of-Domain:**
Do models generalize better to a population of users who have depression but don’t self-report when trained on pre-diagnosis data?
Data

• Self-Report Dataset (Reddit)
  - Based on self-reported diagnosis patterns
  - 20.5K diagnosed users, 9 controls per diagnosed user

• Survey Dataset (Twitter)
  - Based on a survey at the University of Michigan
  - 32 depressed users, 23 with other mental health conditions, 138 controls
Data

• Self-Report Dataset (Reddit)
  • Based on self-reported diagnosis patterns
  • 20.5K diagnosed users, 9 controls per diagnosed user
Data

- **Self-Report Dataset (Reddit)**
  - Based on self-reported diagnosis patterns
  - 20.5K diagnosed users, 9 controls per diagnosed user

- **Survey Dataset (Twitter)**
  - Based on a survey at the University of Michigan
  - 32 depressed users, 23 with other mental health conditions, 138 controls
Data

• Self-Report Dataset (Reddit)
  • Based on self-reported diagnosis patterns
  • 20.5K diagnosed users, 9 controls per diagnosed user

• Survey Dataset (Twitter)
  • Based on a survey at the University of Michigan
  • 32 depressed users, 23 with other mental health conditions
Finding Diagnosis Dates

- Some self-report posts give a hint as to when the user was diagnosed with depression
Finding Diagnosis Dates

- Some self-report posts give a hint as to when the user was diagnosed with depression
- We can determine these dates with **2-week precision for 691 users**
Modeling Setup
Modeling Setup

• **Models**
  • Logistic regression - TF-IDF and LIWC features
  • FastText
  • MentalBERT
Modeling Setup

• **Models**
  • Logistic regression - TF-IDF and LIWC features
  • FastText
  • MentalBERT

We focus on these for brevity – full results in the paper!
Modeling Setup

• **Models**
  - Logistic regression - TF-IDF and LIWC features
  - FastText
  - MentalBERT

• **Training Data Settings**
  - **All-Large**: all data from 20.5K users
  - **Pre-Diagnosis**: data from before diagnosis for 691 users with diagnosis date
  - **All-Small**: All data from the 691 users from Pre-Diagnosis
Modeling Setup

• **Models**
  - Logistic regression - TF-IDF and LIWC features
  - FastText
  - MentalBERT

• **Training Data Settings**
  - **All-Large**: all data from 20.5K users
  - **Pre-Diagnosis**: data from before diagnosis for 691 users with diagnosis date
  - **All-Small**: All data from the 691 users from Pre-Diagnosis
Modeling Setup

- **Models**
  - Logistic regression - TF-IDF and LIWC features
  - FastText
  - MentalBERT

- **Training Data Settings**
  - **All-Large**: all data from 20.5K users
  - **Pre-Diagnosis**: data from before diagnosis for 691 users with diagnosis date
  - **All-Small**: All data from the 691 users from Pre-Diagnosis
Modeling Setup

• **Models**
  - Logistic regression - TF-IDF and LIWC features
  - FastText
  - MentalBERT

• **Training Data Settings**
  - **All-Large**: all data from 20.5K users
  - **Pre-Diagnosis**: data from before diagnosis for 691 users with diagnosis date
  - **All-Small**: All data from the 691 users from Pre-Diagnosis
All-Large Models Outperform Pre-Diagnosis Models on In-Domain Data

![Graph showing F1 scores for different models and data sets]
Pre-Diagnosis Models are Competitive on Out-of-Domain Data (Survey-Based)

The best results overall are with large language models with access to more data.
Pre-Diagnosis Models are Competitive on Out-of-Domain Data (Survey-Based)

With small models, Pre-Diagnosis models are competitive or better than All-Small
All-Large Classifier Weights Reflect Mental Health Discussion

Content warning: explicit text related to suicide appears on the next slide
All-Large Classifier Weights Reflect Mental Health Discussion

pre-diagnosis

All-Large

find his
insecure

his dad my life okay

parents friends

someone better

meds anxious medication lonely

mental health disorder my doctor

kill myself
All-Large Classifier Weights Reflect Mental Health Discussion
All-Large Classifier Weights Reflect Mental Health Discussion
Takeaways

• It is harder for models to classify data that comes from user's pre-diagnosis state.

• Careful data selection can be used to create more generalizable linear models.

• Model weights for pre-diagnosis models correspond more to symptoms while weights for ALL models correspond more to mental health discussion.
Takeaways

• It is harder for models to classify data that comes from user’s pre-diagnosis state
Takeaways

• It is harder for models to classify data that comes from user’s pre-diagnosis state
• Careful data selection can be used to create more generalizable linear models
Takeaways

• It is harder for models to classify data that comes from user’s pre-diagnosis state

• Careful data selection can be used to create more generalizable linear models

• Model weights for pre-diagnosis models correspond more to symptoms while weights for ALL models correspond more to mental health discussion