**Objective:** We created a tool to allow researchers to use Twitter data to identify if a user has a specific disease for recommendation of clinical trial studies. Our software trains a machine learning model on a specific disease to identify Twitter users who potentially have the disease for clinical testing. We used the diabetes to test our model. We identify if the tweeter has diabetes, if another person has diabetes, or if it’s a general statement about diabetes.

**System Architecture**

- **ML Model:** Trains a machine learning model on a specific disease to identify Twitter users who potentially have the disease for clinical testing.
- **MTurk Tool:** Streamlines infrastructure for preparing and uploading tweets to MTurk for labeling.
- **Tweet Scraper:** Collects tweets from twitter.com based on keyword, language.

**Data Collection Funnel**

- **Unfiltered Corpus:** >240,000 Tweets
- **Preprocessing:** ~46,000 With Links
  ~61,000 With No Links
- **MTurk Hit Success (~55%)**
  ~9,500 Tweets
- **Selection Bias (~25%)**
  2400 Final Tweet Corpus

**Current Model Performance**

- **Training Set**
  Corpus size: 2,200
  Avg Accuracy: 72%

- **Test Set**
  Corpus size: 550
  Avg Accuracy: 65%

**Sample Tweet**

Original Tweet:
"@driverRyanReed #DSDshirt is awesome! I'm taking new steps to control my diabetes and joining the movement to Stop Diabetes! #yes"

Tokenized Tweet:
['@', 'driverryanreed', '#', 'dsdshirt', 'is', 'awesome', '!', 'i', 'm', 'taking', 'new', 'steps', 'to', 'control', 'my', 'diabetes', 'and', 'joining', 'the', 'movement', 'to', 'stop', 'diabetes', '#', 'yes']

Sampling of Top Features from Training:
- m taking
- my
- my diabetes
- do you
- get diabetes
- just go

Predicted Label:
Self

**Motivation**

- Penn Hospital researchers require live patients to carry out clinical studies.
- Determining suitable patients without access to medical records is incredibly difficult.
- Manually calling & obtaining permission from enough subjects is incredibly time consuming.
- Streamlined identification system would save time, money, and reduce bias.

**Abstract**

**Project Components**

1. **Data Collection Toolkit**
   - **Tweet Scraper:** Collects tweets from twitter.com based on keyword, language.
2. **SVM Classifier**
   - SciKit-Learn Machine Learning model that trains on ~2,400 labeled tweets and builds an intelligent classifier.
3. **Packaged Prediction Tool**
   - Python scripts enabling researchers to determine users likely to fit disease parameters.

**Languages**

- Language Disambiguation for Disease Analysis
- Identifying Clinical Trial Candidates through Machine Learning

**Authors**

- Nathan Fraenkel
- Varun Gupta
- Jason Lucibello
- Boyang Zhang
- Andy Schwartz

**Faculty Advisors**

- Lyle Ungar
- Andy Schwartz

**Department of Computer & Information Science**

Senior Design Poster Day 2014, University of Pennsylvania