“Can tweets influence Big Pharma?”

Exploring social media with linguistic and machine learning algorithms.

Mike Bartoli, Jay Chok, Brian Cohn, Lakshmi Dharmarajan, Swati Munshaw, Anjali Narayakkadan, Eoin Nugent, Eter Rodriguez

Team Masters Project

A collaboration between Keck Graduate Institute and Eli Lilly and Company.

Research questions

- Is it feasible to establish a methodology, employing analytics, to mine social media to aggregate patient insight?
- If so, how can we best analyze, visualize and learn from the data?

Methods

Data Collection

1. We chose Twitter as our primary data source for its standardized data format, high volume, and its broad consumer reach.
2. Lilly and KGI collaborated with an outside vendor to collect and scrub confidential information from tweets which refer to Alzheimer’s Disease (~66,000 tweets), and Rheumatoid Arthritis(~27,000 tweets) between Mar 2013 and Mar 2014.

Data Manipulation

3. Using a supervised machine learning algorithm (Support Vector Machine), we extracted tweets posted by individuals.
4. Next, we answered informative questions about each tweet using computational (quantitative) and manual (subjective) methods.

Visualization

5. Using open-source software, we generated hierarchichal clusters to understand relatedness of terms. Next we created co-occurence networks based on node centrality in a variety of subsets developed from Data Manipulation #4 using Python and D3.js.

Hierarchichal Clustering

Fig 1. (left) Dendrogram of Alzheimer’s Disease tweets representing individuals categorized as caregivers for at least one person suffering from Alzheimer’s Disease.

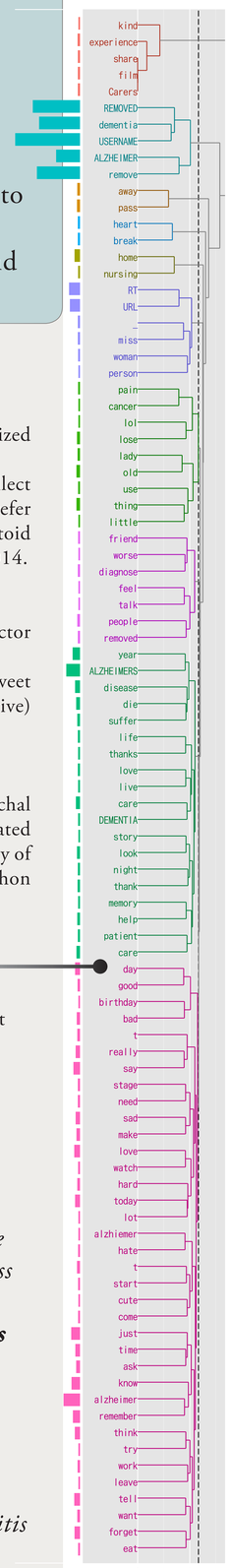
Fig 2. (right) Dendrogram of Rheumatoid Arthritis tweets representing individuals categorized as personally suffering from Arthritis.

Term co-occurrence networks:

For each subset (See Data Manipulation #4), we illustrate how frequently used words appear across tweets. A subsample of each ***linguistic network*** *(in grey) is made into a* ***community-betweeness co-occurence network*** *(in color) to describe associated topics.*

Figures 3-6 : *Networks for Alzheimer’s Disease*
Figures 7-10 *: Networks for Rheumatoid Arthritis*

Figure 1. (below)



Tweets about Alzheimer’s Disease

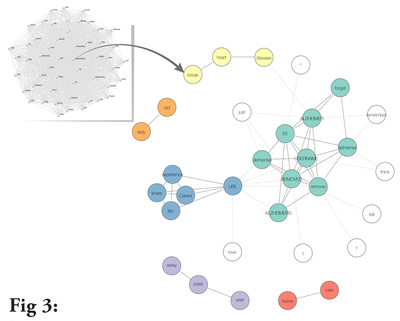


Fig 3: Individuals

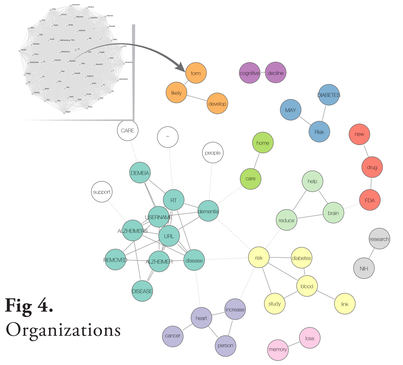


Fig 4. Organizations

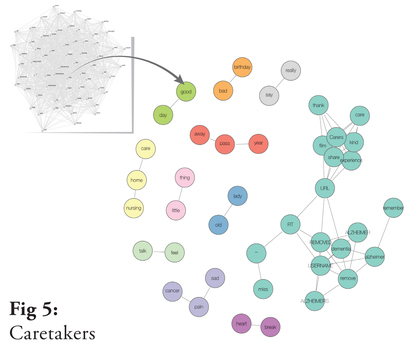


Fig 5: Caretakers

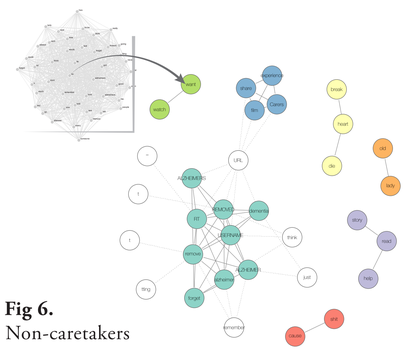


Fig 6. Non-caretakers

Tweets about Rheumatoid Arthritis

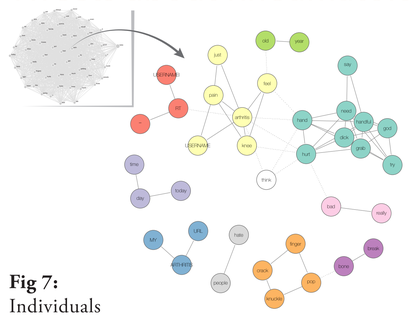


Fig 7: Individuals

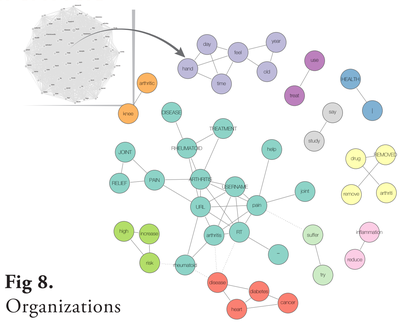


Fig 8. Organizations

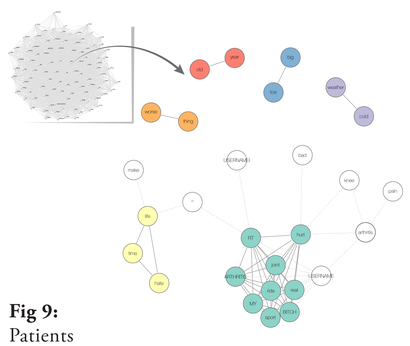


Fig 9: Patients

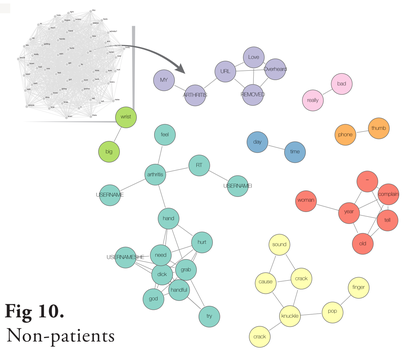
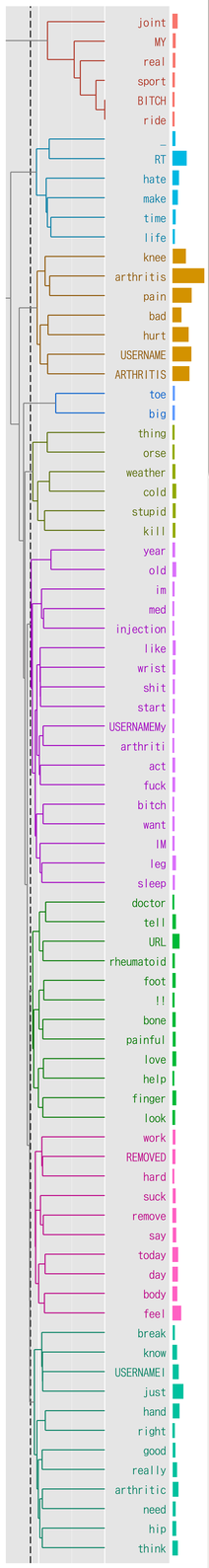


Figure 2. (below)



Discussion

1. FEASIBILITY

Tweets, when mined, may be a source for powerful ‘voice of consumer’ (VOC) data.

2. PROCESS

Mixed methods of computational and manual data mining were most effective at uncovering trends.

This study suggests that Twitter can be a source for patient data in therapeutic areas beyond Alzheimer’s Disease and Rheumatoid Arthritis, and that much of the analyses can be automated for an efficient, cost-effective and powerful tool for social media listening.

Alzheimer’s Disease Caregivers

Post about nursing homes, share their experiences, refer to memories/forgetting, and mourn those who have passed away.

Rheumatoid Arthritis Patients

Use twitter as a space to share their experiences, including their pain, complaints of cold weather, and often use profanity.

Next Steps

- Presenting at the Pharmapack North America 2014 Conference, June ‘14 in New York.
- Exploring more intricate unsupervised machine learning algorithms (including Association Rule Mining and Self Organizing Maps) to uncover deeper nonintuitive trends.

Selected references

1. Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine learning 20(3):273–297.
2. Danowski, J. A., 1993, “Network analysis of message content,” W. D. Richards Jr. & G. A. Barnett eds., Progress in communication sciences IV, Norwood, NJ: Ablex 197-221
3. Fruchterman, T. M. J. & Reingold, E. M. (1991) “Graph Drawing by Force-directed Placement,” Software - Practice and Experience, 21(11):1129-1164.
4. Lampos, V.; De Bie, T.; and Cristianini, N. 2010. Flu detector- tracking epidemics on Twitter. Machine Learning and Knowledge Discovery in Databases 599–602.
5. M Newman and M Girvan (2004) “Finding and evaluating community structure in networks,” Physical Review E 69, 026113
6. Osgood, C.E., 1959, “The Representational Model and Relevant Research Methods,” I. de S. Pool ed., Trends in Content Analysis. Urbana, IL: University of Illinois Press.
7. Signorini, A.; Segre, A.; and Polgreen, P. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PLoS One 6(5).
8. Romesburg, H. C. (1984) Cluster Analysis for Researchers, Belmont, CA: Lifetime Learning Publications