Ontology-based Classification of Social Media Text Data

Sasank Maganti1, Karen A Monsen2, Svetlana Yarosh1
1Department of Computer Science and Engineering, University of Minnesota, Twin Cities 2School of Nursing, University of Minnesota, Twin Cities

INTRODUCTION

- The goal is to classify CaringBridge social media journal entries into various problem concepts using Omaha System taxonomy
- CaringBridge is a web based social network widely used for social support

Health update
Journal entry by check check — 19 minutes ago
I got late to school today as I was not feeling well. I couldn’t stop thinking about my condition and couldn’t sleep last night. I constantly get reminded of the pain I went through and it drives me crazy. I wish I could change that.

Figure 1: Snapshot of a test Journal entry from a CaringBridge site

- Omaha System provides standard taxonomy based terminology for nursing practice

Figure 2: Omaha System

METHODS

- 13,757,900 de-identified CaringBridge journal entries in data corpus
- Preprocessed data with stop words, html text removed and text lemmatized is used
- A 4-step pass on the data to look at the feasibility of using the Omaha system problem concepts
- Related words are derived from consumer versions of the Omaha system, clinical and terminology expertise and internet searches

RESULTS OF FEASIBILITY STUDY

- All 54 problem concepts’ stems were mentioned in journals
- Frequencies of usage ranges from 336 to 2,685,494

Figure 3: Steps in checking feasibility of using the problem concepts

Figure 4: Frequencies of problem concepts after step 2 (red line- combined terms and stems, blue line- step 1)

Figure 5: Frequencies of most frequent problems in journal entries after step 3

Figure 6: Frequencies of most frequent problems in the journal entries after 4 steps

DISCUSSION

- At step 3, 2.5% to 957.3% increase in frequency of problem concept usage
- At step 4, increase by problem concept ranged from 7.4% to 381.2%
- Any problem concept may be found in 11.24% of the journals
- Communicable/infectious condition, spirituality and Neuro-musculo-skeletal function are found in more than 5.8 million journals
- 21 out of 42 problem concepts found in one million or more journals
- Disease specific diagnostic terms were also identified in the data
- Some words like “pray”, “sleep” have very high frequencies that resulted in the respective problem concepts being frequent

FUTURE WORK

- Build models to classify the journal entries into various problem concepts
- Feature selection to train the models
- Features may include problem concept stems, signs and symptoms and related words

ACKNOWLEDGMENT

- This study is funded by CaringBridge, a non-profit organization located in Eagan, Minnesota

References

• http://omahasystem.org/overview.html

Figure 7: Comparison of frequencies of all problem concepts at Step 2 and step 4