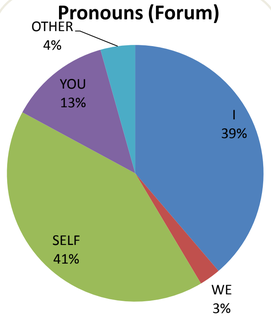
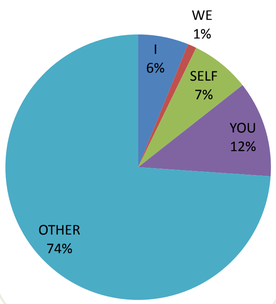
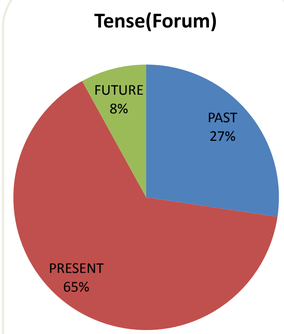
Investigating Text Analysis of User-Generated Contents for Health Related Applications

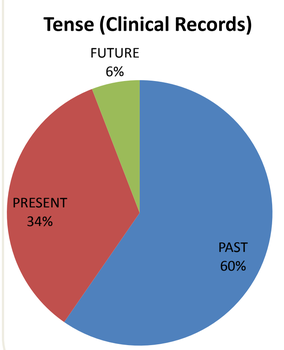
Deema Abdal Hafeth, Amr Ahmed, David Cobham

School of Computer Science, University of Lincoln LN6 7TS Email: dabdalhafeth@lincoln.ac.uk , aahmed@lincoln.ac.uk , dcobham@lincoln.ac.uk









Introduction

Clinical reports includes valuable medical-related information in free-form text which can be extremely useful in aiding/providing better patient care. Text analysis techniques have demonstrated the potential to unlock such information from text. I2b2* designed a smoking challenge requiring the automatic classification of patients in relation to smoking status, based on clinical reports (Uzuner Ö *et al,*2008) . This was motivated by the benefits that such classification and similar extractions can be useful in further studies/research, e.g. asthma studies.

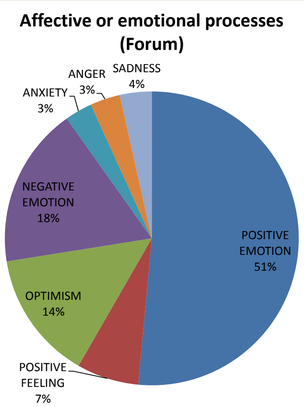
Aim & Motivation

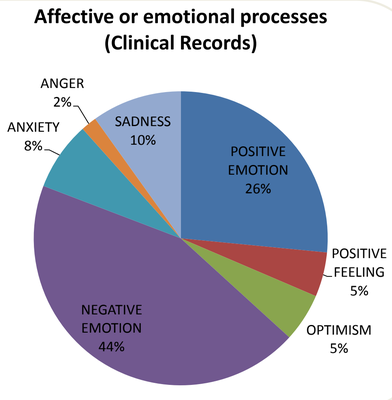
Our aim is to investigate the potential of achieving similar results by analysing the increasing and widely available/accessible online user-generated contents (UGC), e.g. forums. This is motivated by the fact that clinical reports are not widely available and has a long and rigorous process to approve any access.

We also aimed at investigating appropriate compact feature sets that facilitate further level of studies; e.g. Psycholinguistics, as explained later.

Methodology

- Data collected, systematically and with set criteria, from web forums.
- Some properties of the text, for forum data and clinical reports, were extracted to compare the writing style in clinical and forum (shown to the left and below).
- Machine learning (Support Vector Machine) classifier model was built from the collected data, using a baseline feature sets (as per the I2B2 challenge), for each data set (clinical and forum)
- Another model was built using a new feature set LIWC (Linguistic Inquiry and Word Count) + POS (Part of Speech) , for each data set (clinical and forum).
- Smoking status classification accuracy was calculated for each of the above models on each dataset.





*Informatics for Integrating Biology and the Bedside, www.i2b2.org/

Results

- In general, the classification accuracy from forum posts is found to be in line with the baseline results done on clinical records (figure 1).
- Using LIWC+POS features (125 feature) for classification obtained slightly less accuracy, compared to baseline features (>20K feature), but the feature set is compact and facilitates further levels of studies (Psycholinguistic)
- Different factors that affect classification accuracy of forum posts, with LIWC+POS, have been explored, such as (figure 2):
  - Post's length (number of words).
  - Data set size (number of posts).
  - Removing parts of the features .

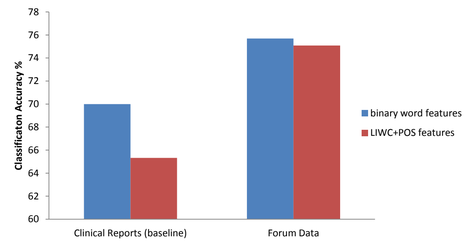


Figure 1. Forum and Clinical Data Classification Accuracy

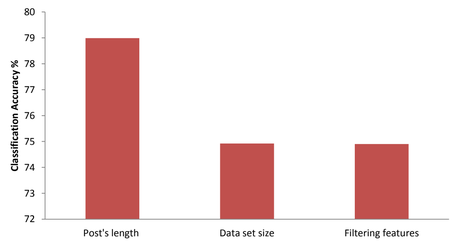


Figure 2. Forum Data Classification Accuracy With LIWC+POS features and Different Factors

Conclusion & Future work

The results suggest that analysing user-generated contents, such as forums, can be as useful as clinical reports. The proposed LIWC+POS feature set, while achieving comparable results, is highly compact and facilitates further levels of studies (e.g. Psycholinguistics).

We expect our work to be for health researchers, medical industrial, by providing them with tools to quantify and better understand people smoking relation and how they behave online, and for forum members, by enriching their use of this rapidly developing and increasingly popular medium by searching for peoples who are in the same situation.

For future work:

• Improve the classification accuracy, with LIWC+POS, and use this feature set as a tool for further and deeper analysis of a person's emotion and psychological status at various stages of the stop smoking process (in journey to stop smoking).

•Integrating other lexical dictionary such as WordNet to capture more colloquiall words and expressions which are not included in LIWC dictionary.

Reference:

Uzuner Ö, Goldstein I, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform. 2008;15(1):14-24. PMID:17947624.