

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.

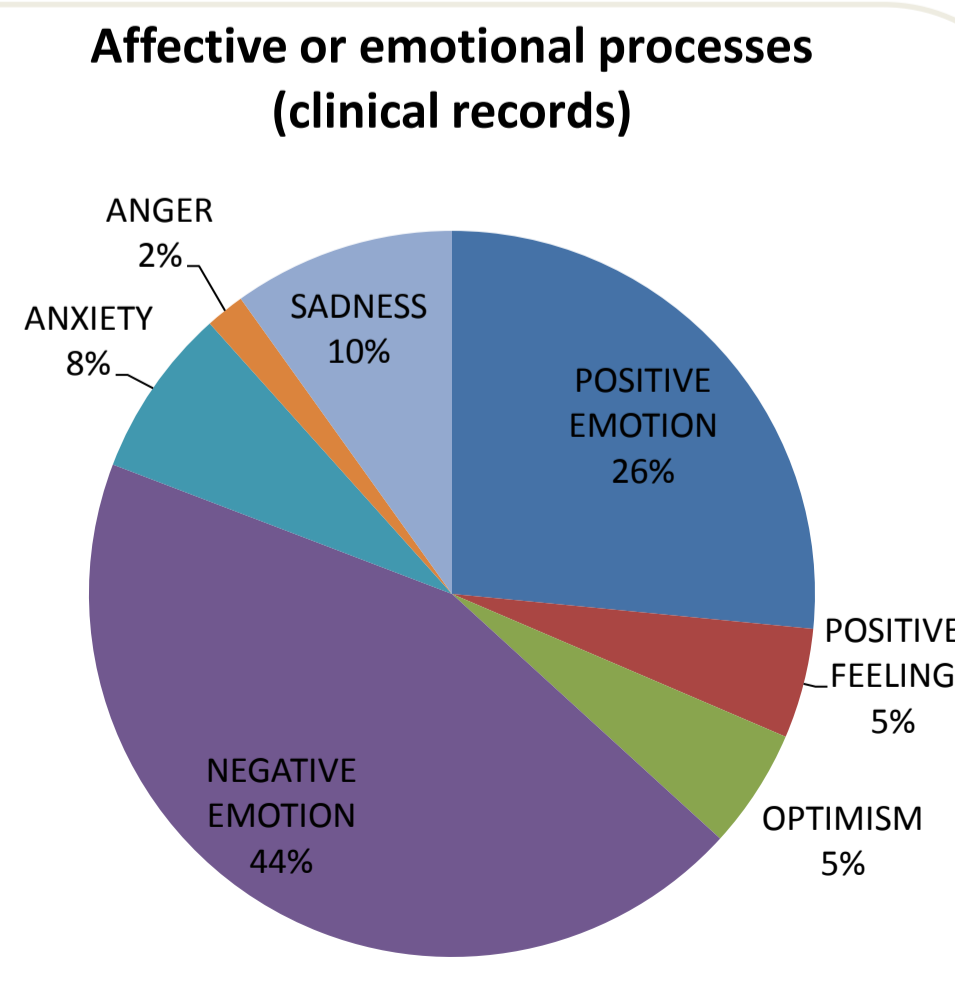
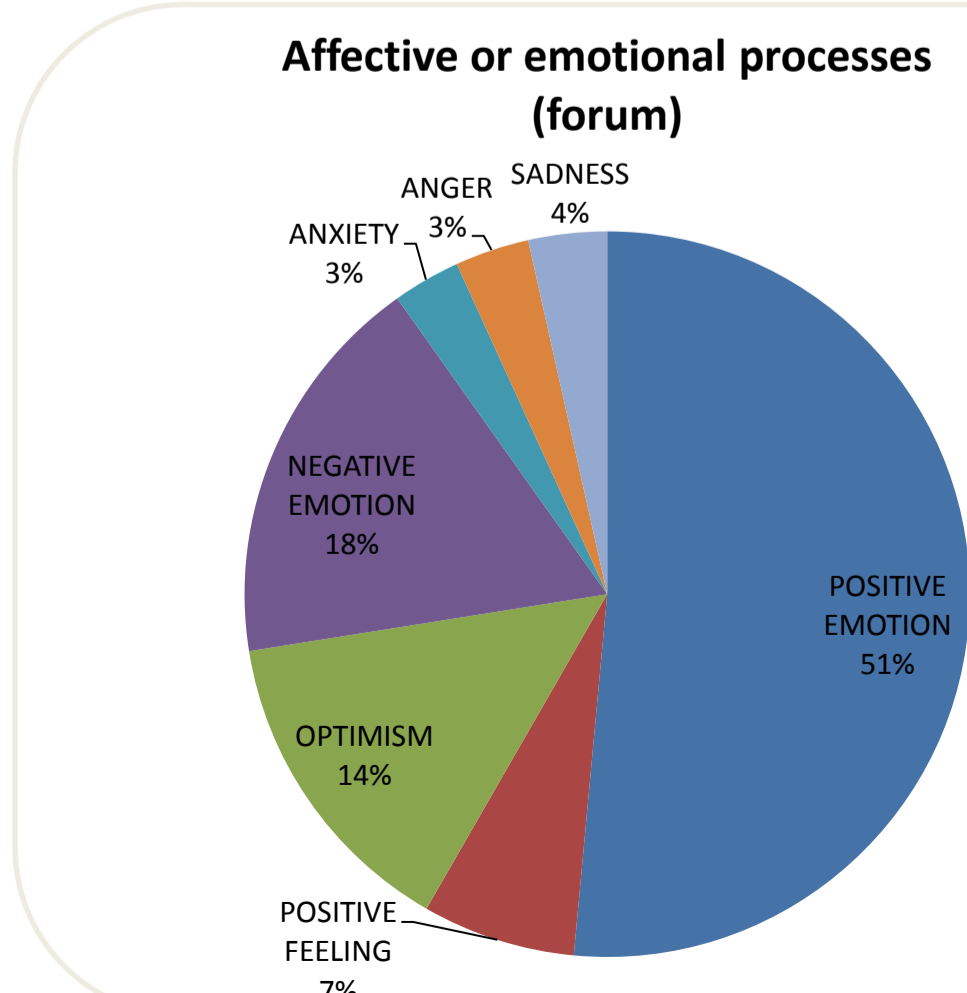
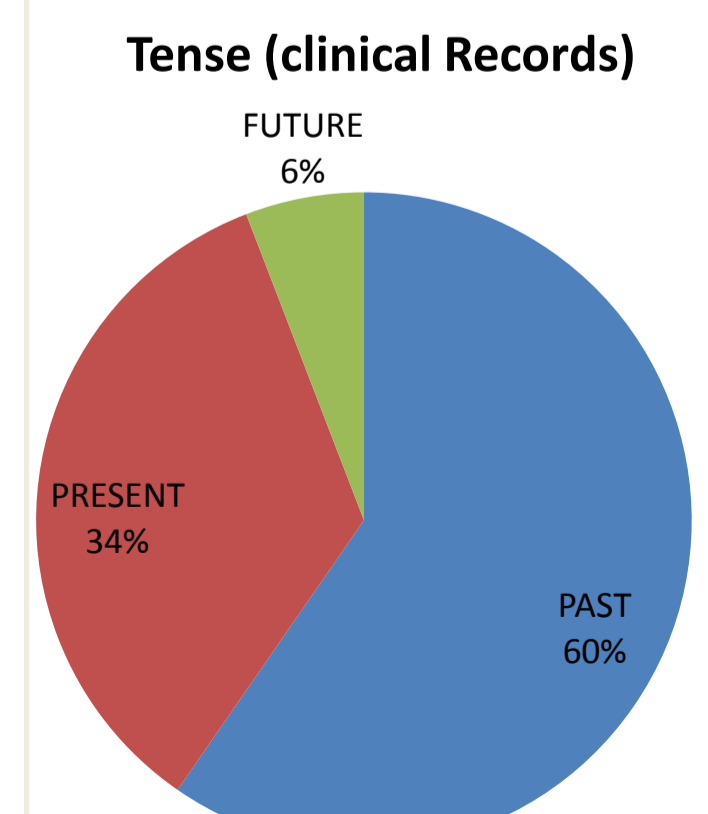
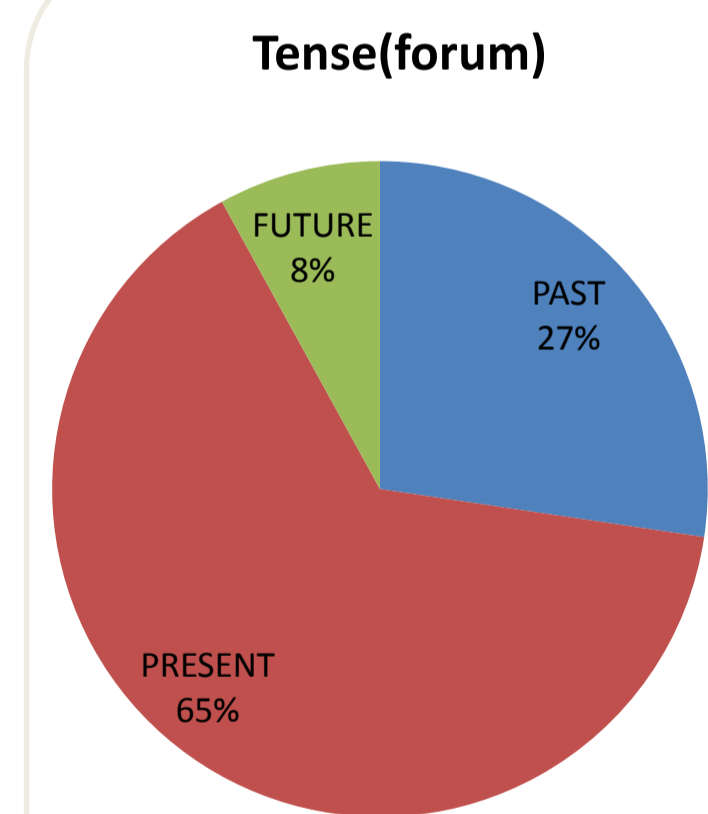
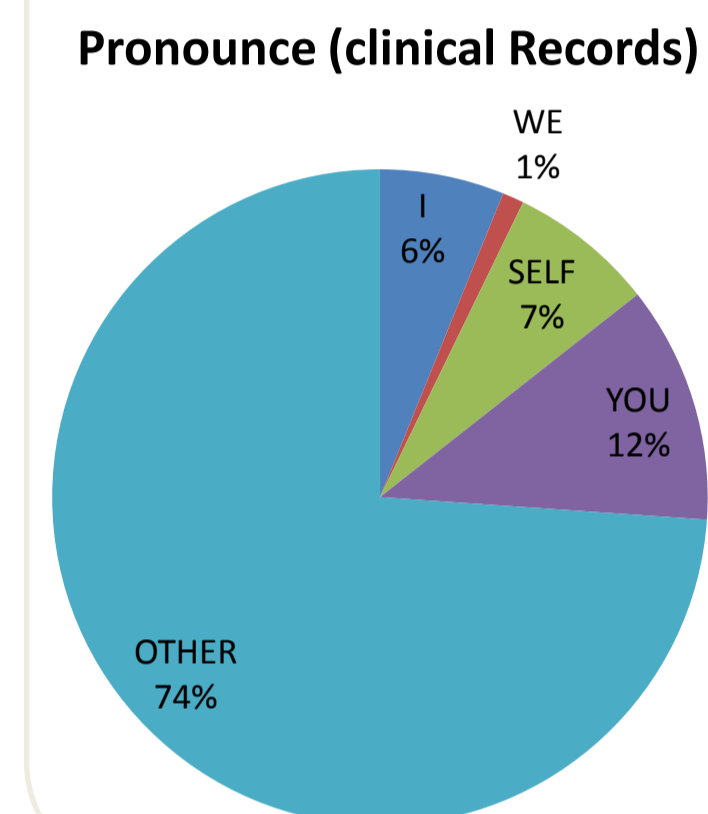
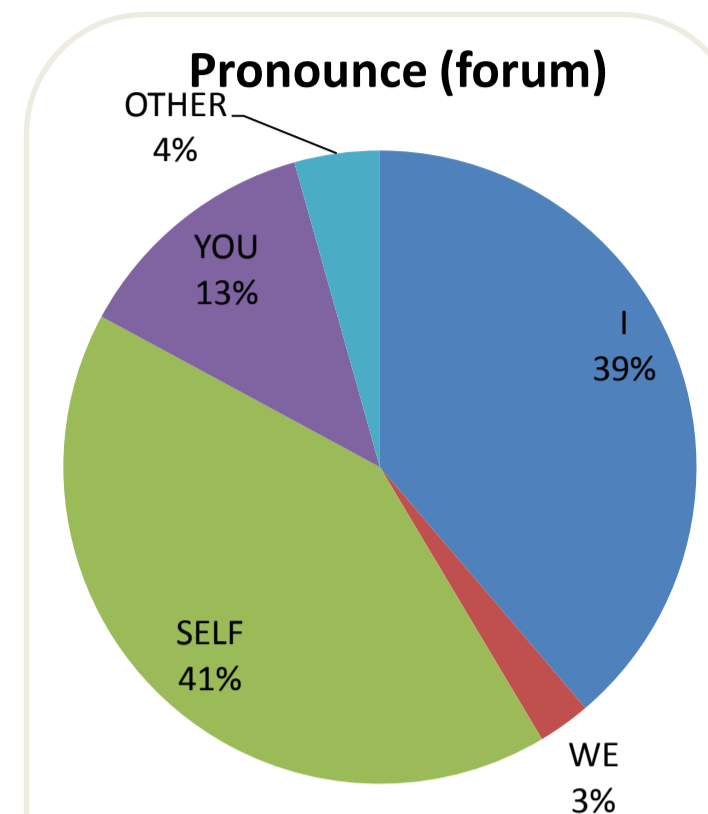
Aim & Motivation

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 two datasets (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.



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) did not improve the accuracy, compared to baseline features (>20K feature). But the feature set is compact and facilitates further levels of studies (Psycholinguistic)
- Classification accuracy of forum posts, with LIWC+POS, can be improved with (figure 2):
 - long post.
 - large data set size.
 - removing parts of the features .

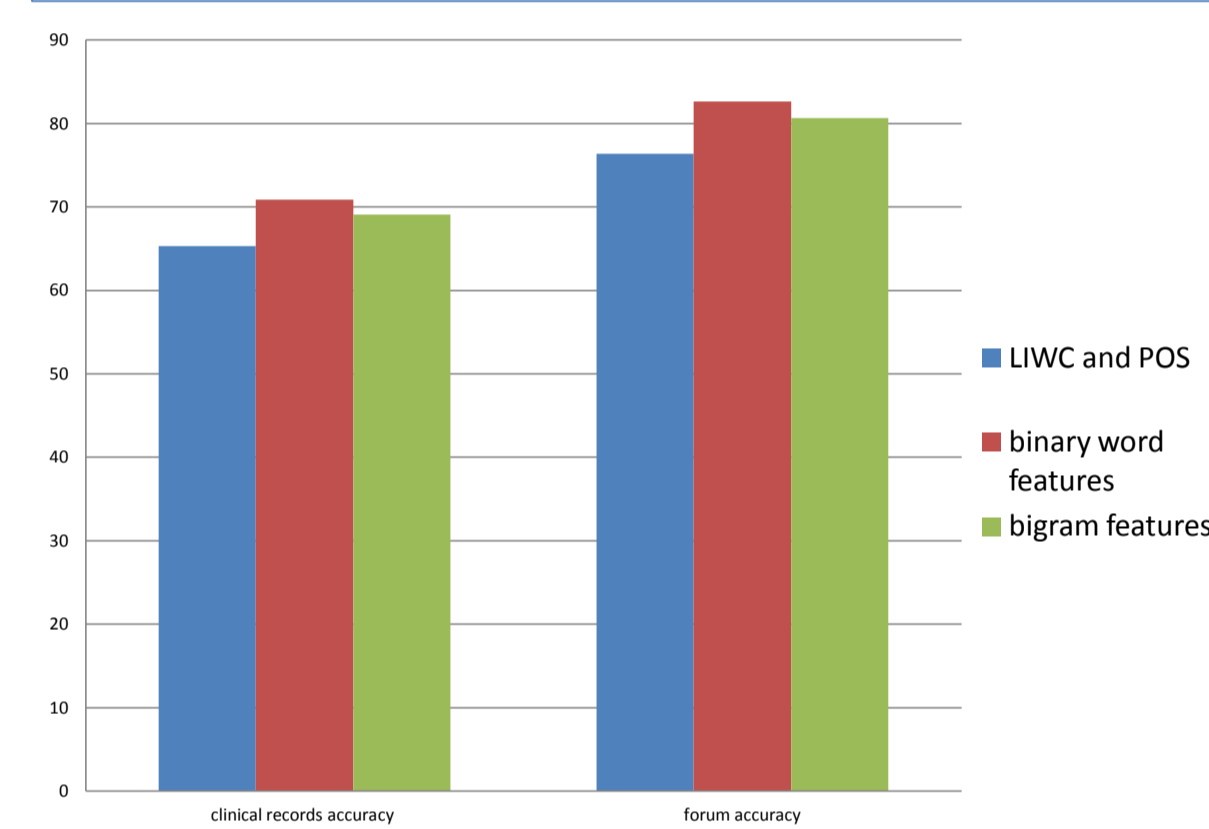


Figure 1. Forum and Clinical Data Classification Accuracy

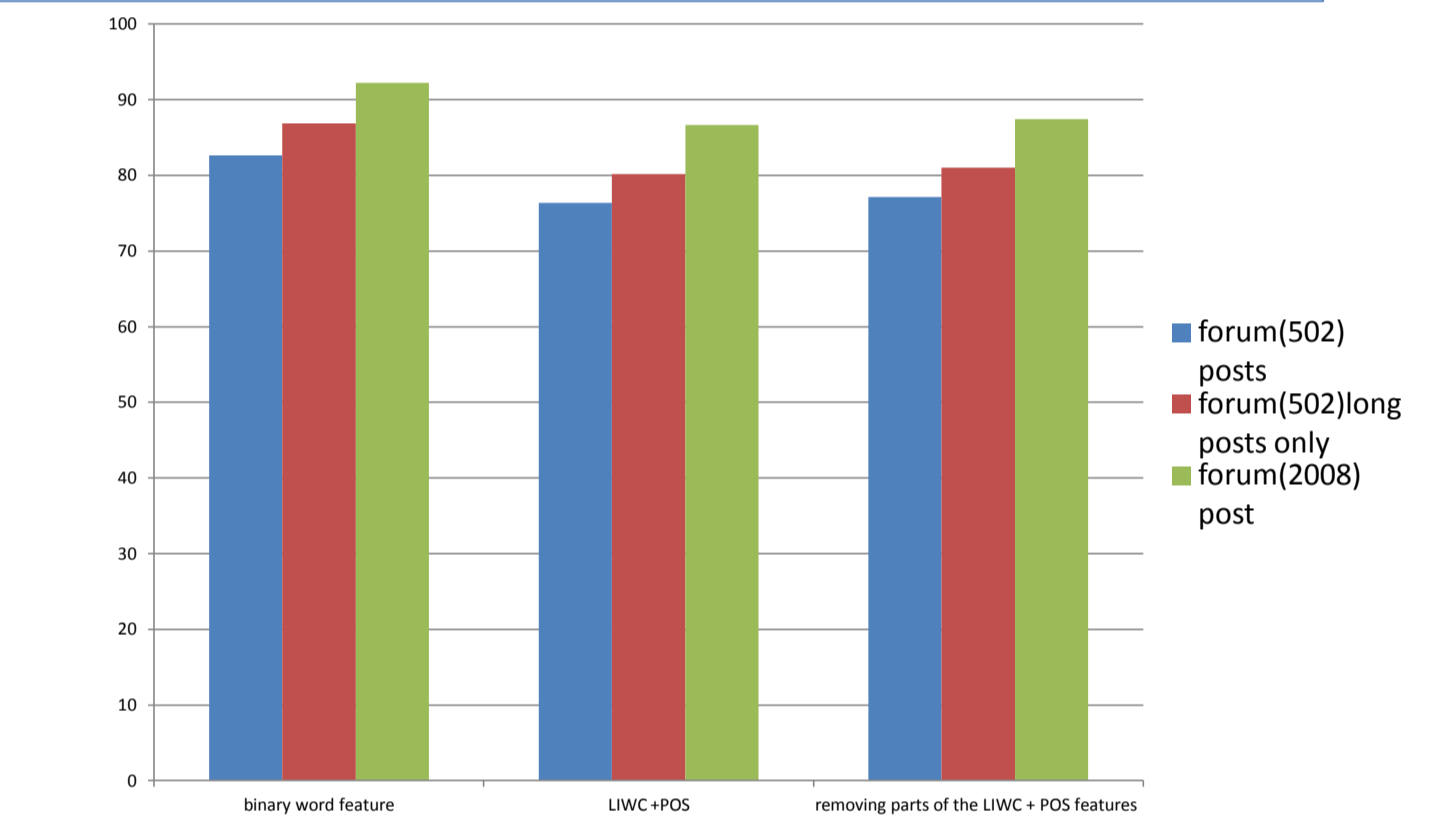


Figure 2. Forum Data Classification Accuracy With Different Features

Conclusion & Future work

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

We expect our work to be useful not only in medical studies but also in Statistical & linguistic studies, access to patient's real-time information, health business (industry)/advertising.

For future work:

- Improve the classification accuracy, with LIWC+POS, and use this feature set as a tool to explore further psychological status and studies.

•Visualisation tool for smokers, in-journey, stop-smoking, past-smoker people to study the process and various factors affecting it, including timings and periods. Similarly the tool could be utilised to identify specific audience (e.g. smokers, in-journey) in forums, to target for specific products or studies.

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.