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

Deema Abdal Hafeth
MSc student by research
School of Computer Science, University of Lincoln

Dr Amr Ahmed
Supervisor

Dr David Cobham
supervisor

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. *This was motivated by the benefits that such classification and similar extractions can be useful in further studies/research, e.g.* ***asthma studies.***

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

Aim and Objectives

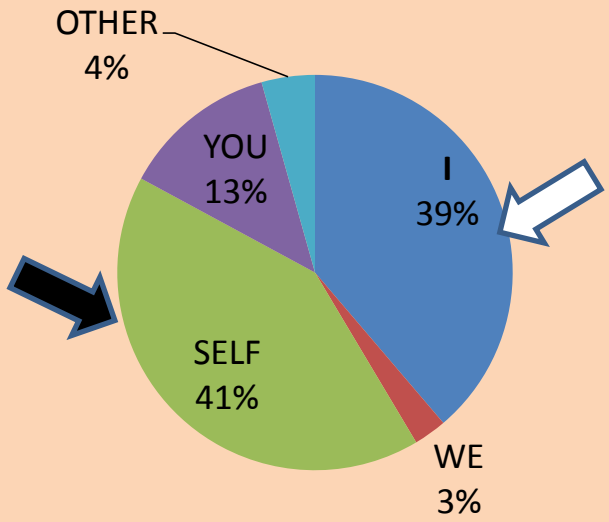
- Investigate the potential of **text analysis techniques** in predicting the **smoking status** but from **user-generated contents such as forums**, in *analogy* with the I2B2 challenge done on the clinical reports.

- Investigate appropriate **compact feature** sets that facilitate further level of studies; e.g. **Psycholinguistics**, as explained later, with the ***hypothesis*** that forum posts have different linguistic features and are rich in personal stories, fresh opinions, and thoughts.

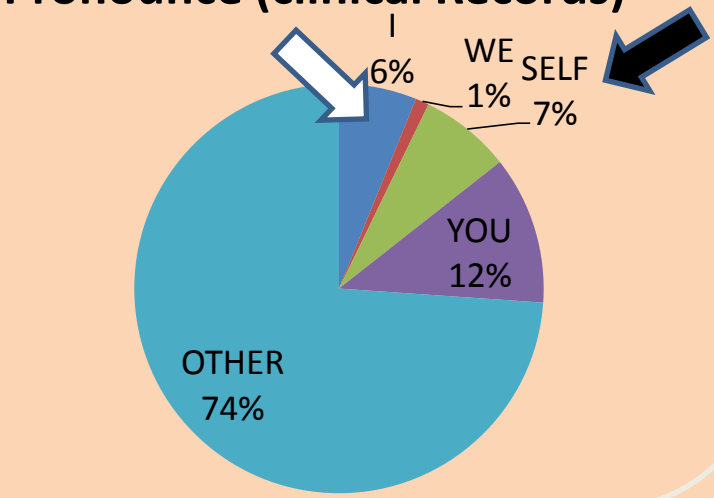
Methodology

- **Data collected**, systematically and with set criteria, from web forums.
- **Extracted and compared** some properties of the text, for forum data and clinical reports.
- 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.

Pronounce (forum)



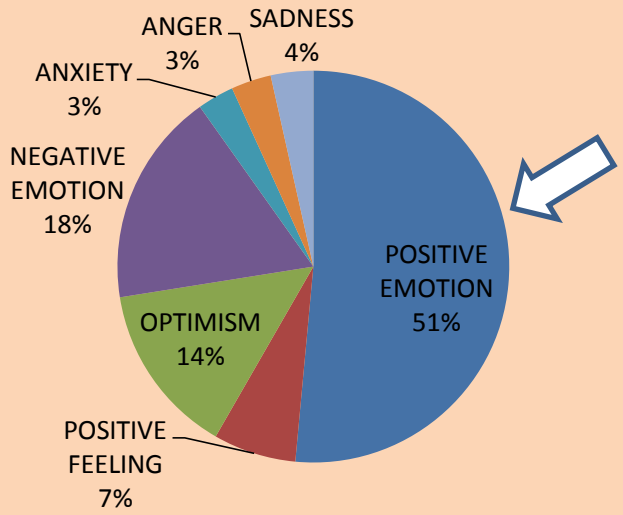
Pronounce (clinical Records)



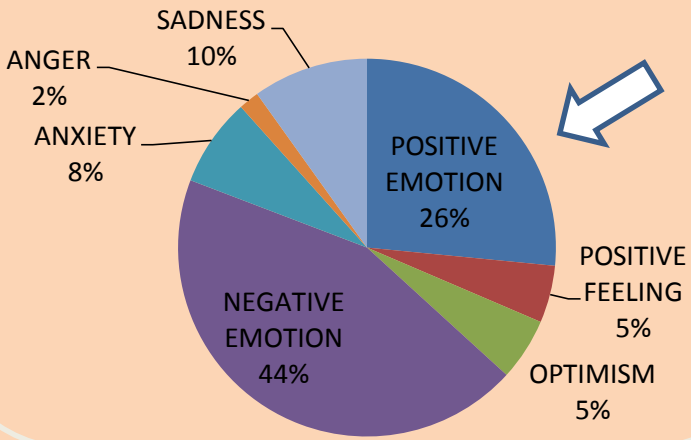
Pronounce

Self	I	Other
I	I	HE
I'D	I'D	HE'D
I'LL	I'LL	HE'LL
I'M	I'M	HE'S
I'VE	I'VE	HER
LET'S	ME	HERS
LETS	MINE	HERSELF
ME	MY	HIM HIMSELF HIS
MINE	MYSELF	SHE
MY		SHE'D
MYSELF		SHE'LL
OUR		SHE'S
OURS		THEIR
OURSELVES		THEM
US		THEMSELVES
WE		THEY
WE'D		THEY'D
WE'LL		THEY'LL
WE'RE		THEY'RE
WE'VE		THEY'VE

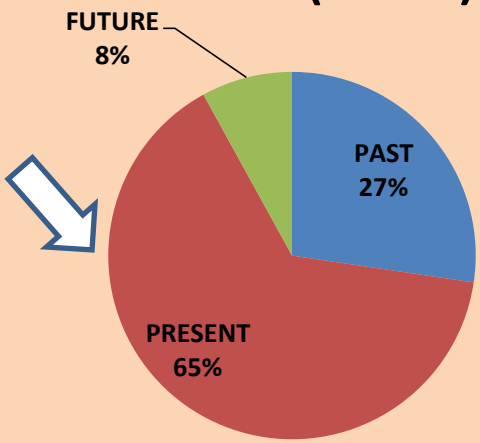
Affective or emotional processes (forum)



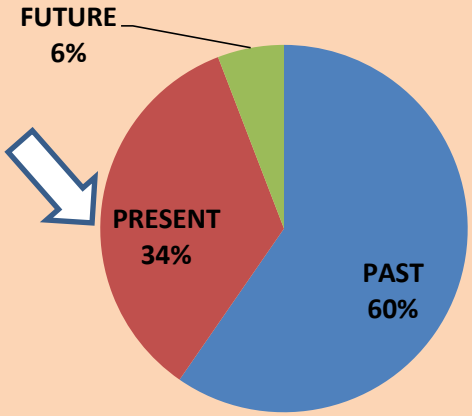
Affective or emotional processes (clinical records)



Tense(forum)

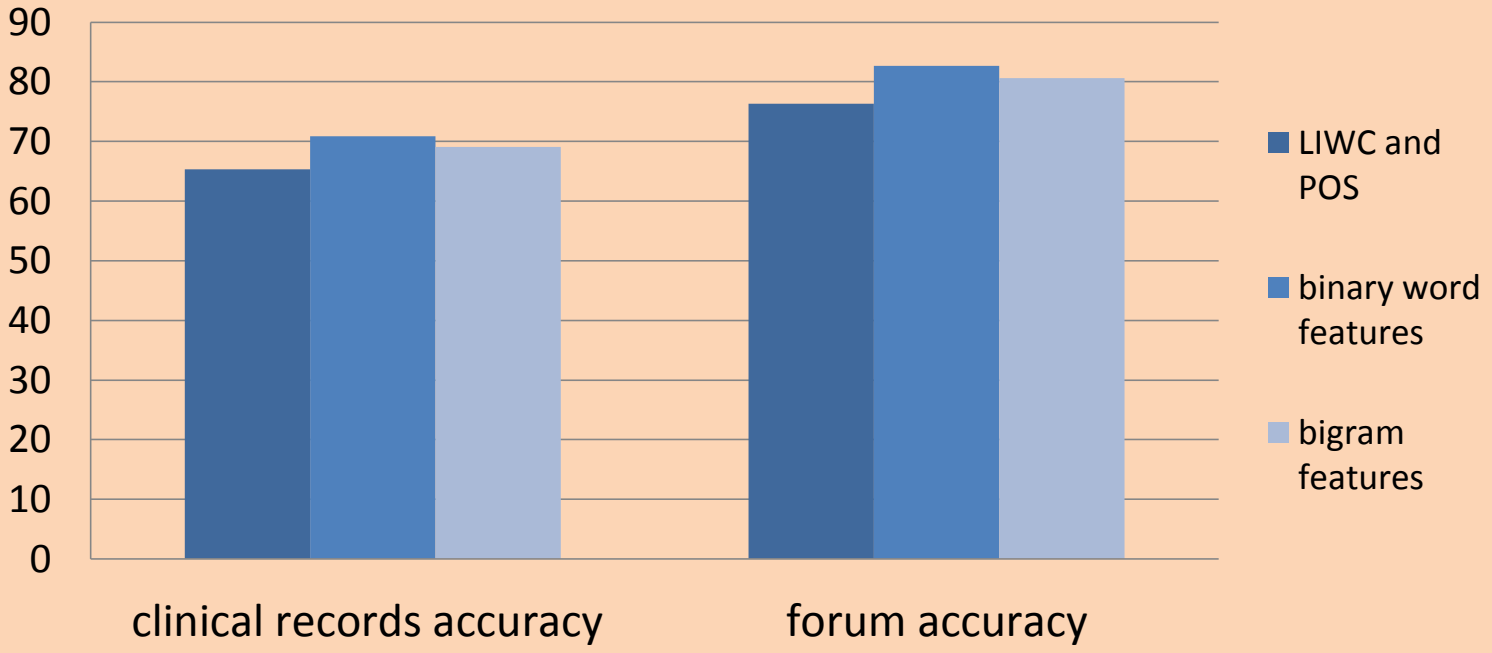


Tense (clinical Records)



Results & Evaluation

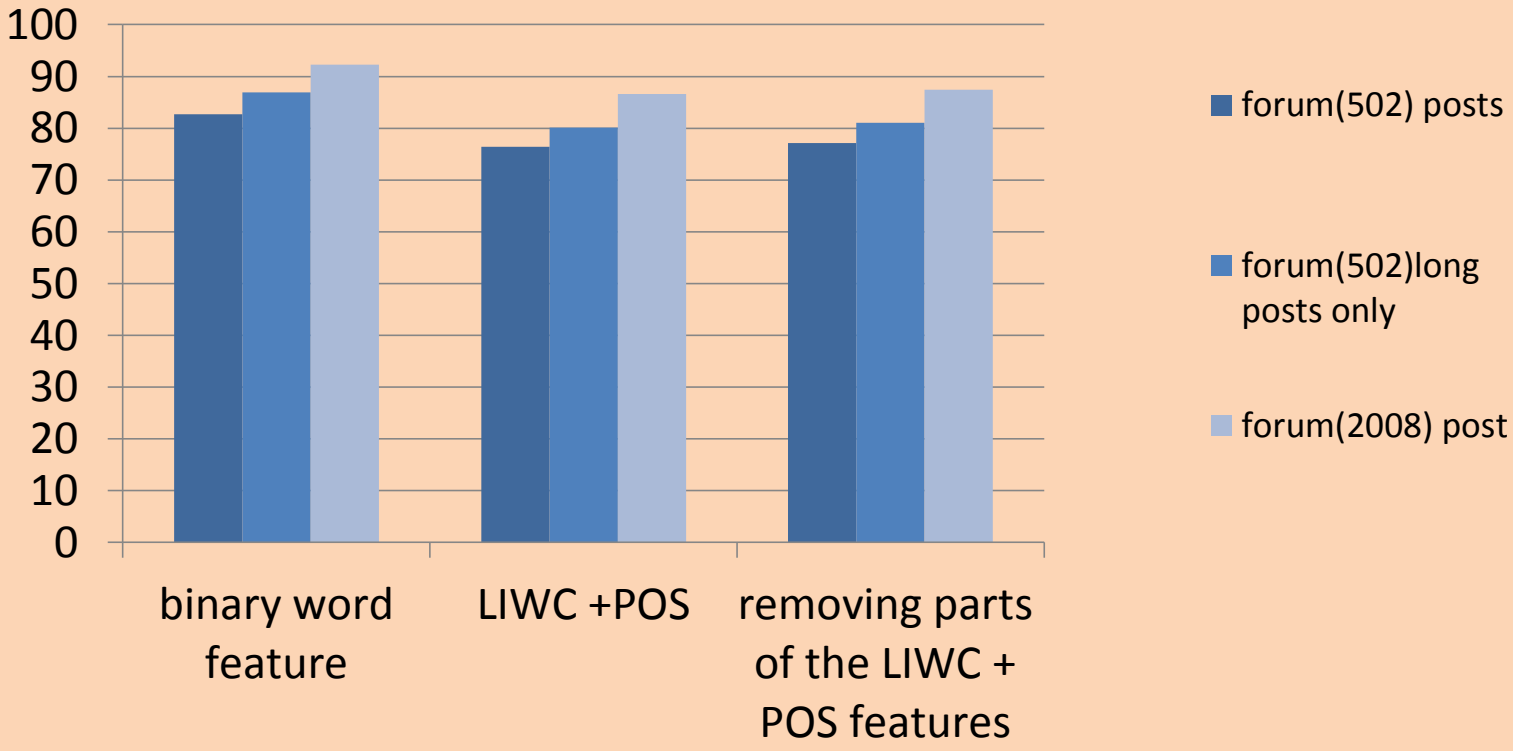
- In general, the classification accuracy from forum posts is found to be in line with the baseline results done on clinical records (figure 1).
- Using **LIWC+POS features (125 feature)** did not improve the accuracy, compared to **baseline features (>20K feature)**. But the feature set is compact, fixed length, independent of the dataset and facilitates further levels of studies (Psycholinguistic)



Results & Evaluation

Forum's classification accuracy with LIWC+POS was improved with :

- **long post**
- **large data set size**
- **removing parts of the features**



Conclusion

•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 to 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.

Thank you

dabdalhafeth@lincoln.ac.uk