Investigating text analysis of user-generated contents for health related applications

Deema AbdalHafeth , Amr Ahmed , David Cobham
dabdalhafeth@lincoln.ac.uk , aahmed@lincoln.ac.uk , dcobham@lincoln.ac.uk

Introduction: Data in patients' records includes free-form text, which have valuable medical related information embedded in. This data can be extremely useful in aiding and providing better patient care. Text analysis techniques have demonstrated the potential to unlock such information from text. For example, the I2B2 (the Information for Integrating Biology and the Bedside) designed a smoking challenge that required the automatic classification of patients in relation to smoking status, based on clinical reports. The challenge focused on text analysis as a powerful tool to classify clinical records and detect a patient's smoking status. Determining smoking status, from the text, is beneficial for further studies and research, such as asthma studies.

Motivation: One challenge with clinical reports' data is their strict availability and difficulties in accessing them, including the long process of approvals required. On the other hand, people are expressing themselves more widely nowadays and the online user-generated contents (UGC), like forums and blogs, are becoming more available. Those user-generated contents are written by people themselves and are usually publicly available, which make them easily accessible. If analysing this available and accessible data, instead of the clinical records that are difficult to access, can result/reveal similar information, this would definitely help. The availability and accessibility of the data will also motivate more researchers to work on it, which will results in more advances in the research and benefits.

Aim and Objectives: The aim of this work is to 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. This especially includes the use of Psycholinguistic features on analysing forms, with the hypothesis that forum posts have different linguistic features and are rich in personal stories, fresh opinions, and thoughts.

Methodology:

First, we explored the differences between the properties of each data type as clinical reports (figure 1) are written by clinicians, while forum posts (figure 2) are written by the users themselves who are usually ordinary people (not specialist). Then, we investigated the relevant features and techniques to achieve the task of classifying forum posts into 4 classes; *non-smoker*, *current smoker, in-journey to stop smoking* and *past smoker*.

Data collection: Forum data was collected from forum websites by using Google's search engine and using four different kinds of queries to find relevant posts (in-journey, past smoker, non-smoker and current smoker). To obtain enough and covering data, a set of criterion was systematically applied on the collected data such as availability of posts for the in-journey duration, frequency of posting, status indicated by the user, …etc. Clinical records data was obtained from I2B2.

Features: A number of feature sets were investigated, some from the previous work in I2B2 challenge (e.g. Uni-gram, Bi-gram, Word frequency, TF/IDF). But we also investigated the use of Psycholinguistic features, such as LIWC (Linguistic Inquiry and Word Count) and POS (Part of Speech). This is also one of the contributions in this work. We investigated each of those features on both types of data; forum posts and clinical reports, and compared the results through the classification task.

Classification: The extracted features were fed into machine learning algorithms to build a classifier for predicting the smoking status. We experimented with several machine learning algorithms and various factors that may affect/improve classification accuracy, such as the post length, size of the training data set, arbitrarily and systematically removing parts of the features.

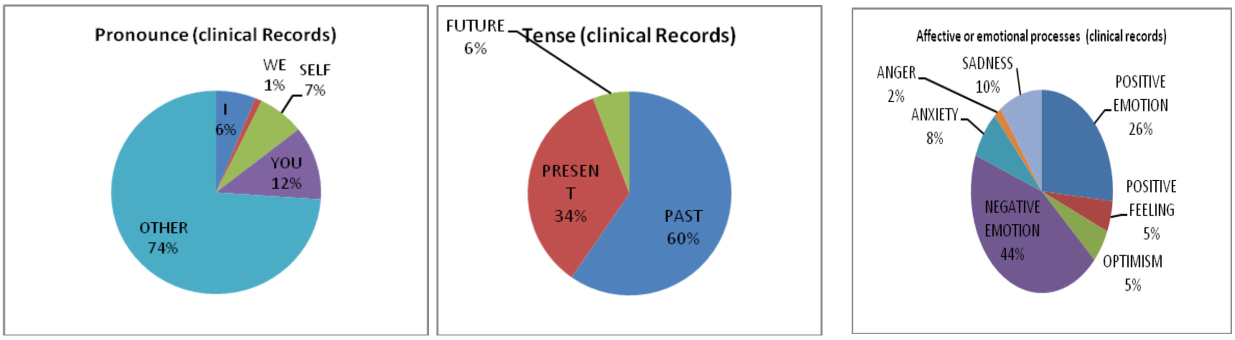

Figure 1: Properties of the *Clinical Records* dataset

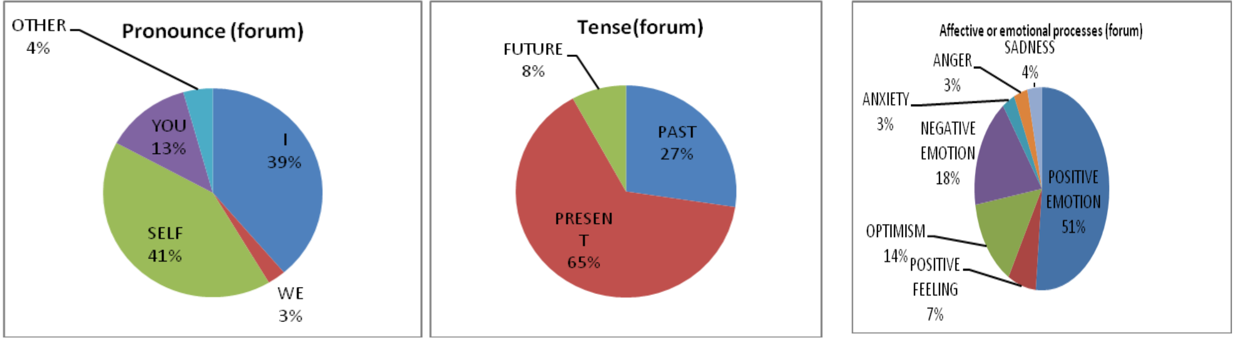

Figure 2: Properties of the *Forums* dataset

Results & Evaluation: The obtained classification results are compared to other methods documented in the literature, mainly with the I2B2 challenge for clinical reports. The classification accuracy, on the forum posts, was found to be in line with other experiments in the literature, on the clinical reports (figure 3). This suggests that forum data could become an important resource for the task of smoking status classification, and similar tasks.

Regarding the features, it was found that the uni-gram features are giving the highest accuracy in both data sets, with the LIWC+POS having a slightly less accuracy. However, the size of the feature vector varies with the dataset and could become huge (20K+). When it was cut down, the accuracy decreased. On the other hand, using the LIWC and POS, although have a slightly less accuracy, their feature vector is relatively highly compact (only 125). More importantly, this compact feature vector is of fixed length, independent of the dataset. Moreover, the LIWC contains Psycholinguistic categories, such as emotion, affective, …, etc. Those categories are promising in further and deeper analysis of the person's emotion and psychological status at various stages of the stop-smoking process (or similar tasks). This suggests that LIWC+POS are potential candidate features set for such tasks, and could be a potential tool to explore further psychological status and studies, in addition to this classification task.

The above results are found to be promising, both in terms of the forum data (which is easier to obtain, compared to clinical records) and for the compact feature set. Various applications are envisaged, based on the above. Examples include visualisation tool for smokers, in-journey, stop-smoking, past-smoker people to study the process and various factors affecting it, including timings and periods. Similar application could be identifying specific audience (e.g. smokers, in-journey) in forums, to target for specific products or studies.

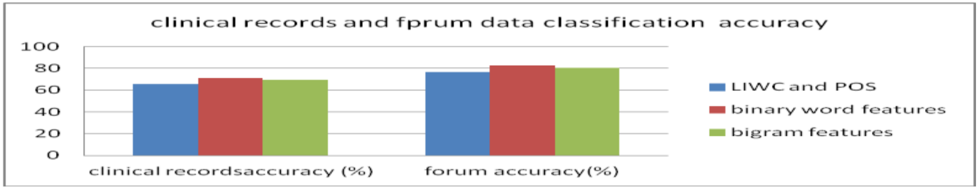

Figure 3: Classification results, on both datasets, with our features set compared to baselines

Potential future work: The approach can be applied in a visionary-related text (e.g. consultation and diagnosis report), linked to image features (e.g. images of rashes) and categorising them in general or more specifically to classify specific skin diseases such as eczema or psoriasis.