Text Analysis of User-Generated Contents for Health-Care Applications: Case Study on Smoking Status Classification

Deema Abdal Hafeth, Amr Ahmed, David Cobham
School of Computer Science, University of Lincoln, UK
{dabdalhafeth, aahmed, dcobham}@lincoln.ac.uk

Keywords: Smoking status classification, Text mining, User-Generated Contents.

Abstract: Text mining techniques have demonstrated a potential to unlock significant patient health information from unstructured text. However, most of the published work has been done using clinical reports, which are difficult to access due to patient confidentiality. In this paper, we present an investigation of text analysis for smoking status classification from User-Generated Contents (UGC), such as online forum discussions. UGC are more widely available, compared to clinical reports. Based on analyzing the properties of UGC, we propose the use of Linguistic Inquiry Word Count (LIWC) an approach being used for the first time for such a health-related task. We also explore various factors that affect the classification performance. The experimental results and evaluation indicate that the forum classification performs well with the proposed features. It has achieved an accuracy of up to 75% for smoking status prediction. Furthermore, the utilized features set is compact (88 features only) and independent of the dataset size.

1 INTRODUCTION

The increasing availability of text collections allows researchers to apply text classification for predictive purposes relating to topics, opinions, moods, diseases and personalities. One of the active research areas in text classification is that of smoking status classification. Smoking status identification is the process of discovering, or distinguishing, the smoking status of the author of a given particular text from a set of predefined categories. Interest in automatic smoking status classification started in 2005 particularly for clinical records. This type of data is difficult to obtain, due to the lengthy approvals process that protect patient confidentiality. In addition, prior to this study no specific features have been identified as being a standard for such a classification. These issues pose additional challenges for a researcher undertaking further studies using this more traditional source of data.

Web 2.0 has facilitated many forms of interactive collaboration mediated over the internet. In addition, people consider themselves to be more informed and empowered and find it a supportive environment as they can exchange information and experiences from others in the same situation and receive emotional support from them. This new medium is also beneficial for health professionals, as it offers exciting new research avenues with regard to theories of psycho-social support and how people manage their conditions. In this paper, the source data is derived from online forum discussions a source that is more widely and readily available than other types of text such as clinical reports.

In this paper, we present an investigation of smoking status classification from UGC, and introduce more relevant feature sets. The technique used in this paper utilizes the LIWC dictionary (Linguistic inquiry and word count, October 2013). This selection is based on the properties observed in the UGC data, as discussed in Section 3.

The rest of the paper is organized as follows: Section 2 presents the key work relevant to this investigation, The properties of forums are studied and discussed in Section 3, then the proposed features set is introduced in Section 4. The proposed classification framework is discussed in Section 5, followed by the experimentation setup, results and evaluation in Section 6. Finally, the paper is concluded in Section 7.

2 RELATED WORKS

Automated classification of documents is one of the common tasks in text analysis and Natural Language Processing. Examples of applications include opinions classification (Kaiser and Bodendorf, 2012; Pang, Lee, and Vaithyanathan, 2002), and mood inference (Leshed and Kaye, 2006). In the smoking status inferring field, the i2b2 (the Informatics for Integrating Biology and the Bedside) challenge was designed and facilitated by the National Centre for Biomedical Computing (Informatics for integrating biology and the bedside, October 2013). The challenge required participants to explore text analysis for the automatic classification of patients in relation to their 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. Different methods were proposed to achieve this task (Uzuner et al., 2008). We review the work that is most closely related to our research.

The earliest smoking status classification research was done by Sordo and Zeng (Sordo and Zeng, 2005), when they classified clinical reports to find the effect of training set size on classification results. They found that the size of the training set and the classification accuracy are in fact correlated. This method classified clinical reports into one of four smoking statuses (current smoker, past smoker, never smoker, denies smoking) by using Support Vector Machine (SVM), Naïve Bayes (NB) and other machine learning classifiers. In this research, word frequency of unigram and bi-gram were used as features. The approach achieved results of considerable accuracy (86.8%) and concluded that the SVM algorithm is more suitable for classifying larger corpora for smoking status classification.

Clark et al (Clark et al., 2008) developed the best performing system in the i2b2 challenge. They used the binary presence of unigram and bigram word features of document with SVM classifier algorithm and 10 cross-validation techniques. This method achieved an accuracy of 82%. The system performance was then improved by using additional clinical report data and filtering unrelated smoking sentences before classification.

Supervised and unsupervised methods were suggested by Pedersen (Pedersen, 2006) for predicting the smoking status of patients from clinical reports. This involved testing a number of learning classifiers for supervised approaches such as SVM and NB. Furthermore, the frequencies of unigram, bigram and trigram words that appeared in the training set were used as features. The classification method was achieved accuracy of 82%.

In this research, we differ from previous methods in the type of data (UGC) and the relevant features to be used in the smoking status prediction. This is expanded upon in Sections3 and 4.

On the other hand, a rule based system to infer patient smoking status was developed (Wicentowski and Sydes, 2008). It removed all smoking related terms from the training set and created a "smoke blind" set by depending on general information in the document for predicting a smoking label. In this system, a NB classifier algorithm was applied with the bigram word features. The shared objective of this work with ours is the dependency upon general text information for smoking status classification, instead of smoke-related information.

Other approaches have made good progress towards extracting and classifying sentences related to smoking only, and ignoring others as noisy data, by utilising rule-based systems (Liu et al., 2012; Savova et al., 2008; Zeng et al., 2006; Aramaki et al., 2006; Szarvas et al., 2006; Cohen, 2008).

Unfortunately, these methods used in these previous studies, in terms of extracting sentences related to smoking only and classifying each one individually, are not suitable for online forum discussion data. There are two reasons behind that. First, a forum is written by the users themselves and sentence boundaries are not guaranteed due to poor use of punctuation. Second, forums have fewer words and a smaller number of sentences per post in comparison with clinical reports.

All the above methods designed to predict smoking status, share a common data type, namely the clinical reports corpus. These methods are may not be directly applicable to forum posts due to the different nature of the text in clinical reports. In addition to that, clinical reports have limited availability and are difficult to access. Furthermore, it can be seen that smoking status classification utilising online forum discussion has received little attention from researchers. Consequently, no specific standard features have been confirmed or recommended for smoking status classification in previous literature.

In this work, we address the above issues by investigating smoking status classification on a different type of data source and identify relevant features to use in smoking status classification. The main focus is on UGC data such as online forum discussions. In the next section, the comparison in writing style and text properties for online forum discussion and clinical reports is presented.

3 FORUM LANGUAGE PROPERTIES

The style of online forum discussion is different to other types of texts, such as clinical reports. The nature and the properties of the language in forum and clinical reports are presented in this section.

The text in forums is less focused and directed than that in clinical reports. It contains thoughts, everyday experiences, feelings, opinions, and social status. Furthermore, as it is written by the users themselves, it is less grammatically and syntactically accurate than clinical reports. It enjoys almost universal public access, with no pre-determination involved in terms of criteria for specific readers. The text has the advantage of being written in colloquial language and unedited. This complexity of text motivates us to search for the best features that capture smoking status. Furthermore, The styles of writing are able to reflect the person's situation and are useful in research avenues like person's personality, emotions and social state.

A summary of the language properties, of both forums and clinical reports, are shown in Figure 1. A high percentage of usage of first person singular pronouns, positive emotion and the present tense concur with the forum corpus, in contrast with clinical reports. First person singular pronouns hold a dominant position in the poster's writing, because this data is written by the users themselves. Likewise as the events or everyday activities are immediately reported, the present tense is used widely. Various subjects in addition to health related issues are also covered due to the authors feeling free to include them in their posts. As a result, more words expressing positive emotion are used rather than words that express negative emotion. Those characteristics require new types of features that can help determine the smoking status class. In the following section the selected features for this investigation are explained in detail.

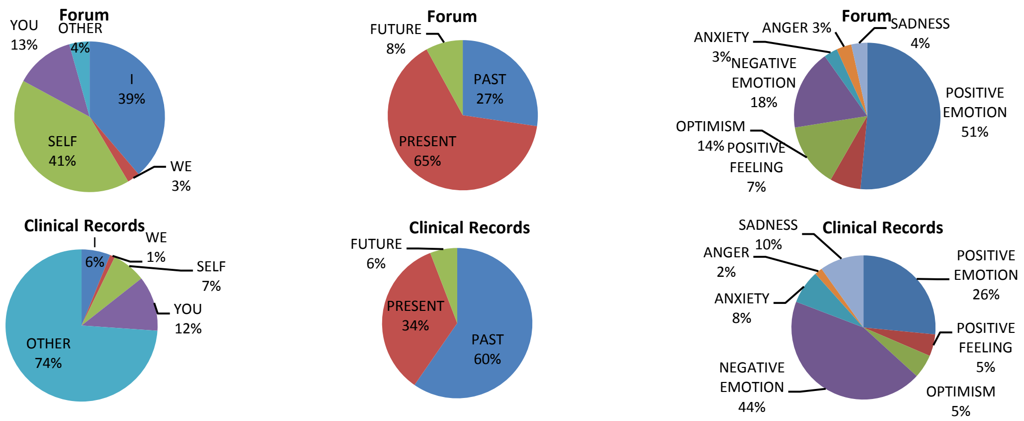
Furthermore, to explore the feasibility of applying this proposed approach to other types of UGC, language properties for blog, email and online forum texts were compared. To the best of our knowledge, the comparison in linguistic style, for these types of UGC, has not been investigated.

A group of fifteen psycholinguistic features for blog and email have been produced by (Gill, Nowson, and Oberlander, 2006), more features were extracted from linguistic and psychological main categories. The features belonging to blog and email were compared with the same features in forum and clinical records. The comparison shows that email and blog features are almost in line forum data but differ from clinical reports.

In conclusion, this approach of smoking status classification could be generalised on data of UGC sources, as they share common psycholinguistic properties.

4 THE PROPOSED FEATURE SET

In text classification task, the features selection has a crucial role to play in the final results in text classification tasks. In the course of analyzing the nature of the online forum's text, in Section 3 we have demonstrated that the psycholinguistic features of writing forums are different from other sources of text (clinical reports). Therefore, psycholinguistic features



Pronounce Tense Emotional processes

Figure 1: Psycholinguistic features in clinical reports and forum data corpora.

are proposed to apply in this research to represent feelings, personal activities and thoughts that are included in forum text.

LIWC dictionary has been selected as a feature set for smoking status classification. LIWC counts the appearance of words or word-stems belonging to pre-defined psychological and linguistic categories. For example, the term “hunger” captures the words hungry, hungrier, hungriest. Furthermore, one of the major strengths of the LIWC is that the dictionary has been rated and evaluated by independent judges (Tausczik and Pennebaker, 2010).

The selected 88 LIWC features were grouped into four types:

- Standard linguistic features (e.g., total word count, pronouns);
- Psychological features (e.g., cognitive, emotional processes);
- Personal concerns features (e.g., occupation, leisure activity);
- Paralinguistic features (e.g., non-fluencies, fillers words).

5 THE SMOKING STATUS CLASSIFICATION FRAMEWORK

The methodology of evaluating online forum discussion for smoking status classification involved executing 3 experiments. The aim of experiment 1 was to evaluate the results of classifying forum data with baseline results (that was dependent on clinical reports). In the second experiment, the performance of using a LIWC features set was compared with other features type. The aim of experiment 3 was to explore the other factors that effected the classification results.

Three fundamental steps were conducted as a part of smoking status classification framework, as depicted in Figure 2. These steps involved:

- Pre-processing phase: includes obtaining the data from the web and cleaning them. For example, removing repeated and empty posts;
- Feature Extraction phase: converts each posting into a corresponding features vector. This changes the input data from unstructured text space into features vector space;
- Model building phase: includes the use of a suitable machine learning algorithm classifier that produces a useful classification model. The extracted model is then tested and evaluated in terms of having the best results.

6 EXPEREMENTS, RESULTS, AND DISCUSSION

In this section the experiments' setup, datasets, results and evaluations are detailed.

6.1 Experiment Setup

To evaluate the proposed framework, various experiments were carried out over online forum discussion corpus and clinical reports datasets. The evaluations included comparison with baseline method, applying different types of features and exploring factors that affect the framework results.

The experiments have been performed using the SVM machine learning algorithm, as it is one of the best algorithms available in data mining tasks (Wu et al., 2008). A 10-fold cross validation procedure was used in training and testing each dataset by using WEKA toolbox (The University of Weka to October 2013). The experiments ran on Intel Core i5- 2.30 GHZ computer with 6 Gigabytes of RAM.

6.2 Datasets

The experiments have been performed using various forum discussions datasets, a collection of

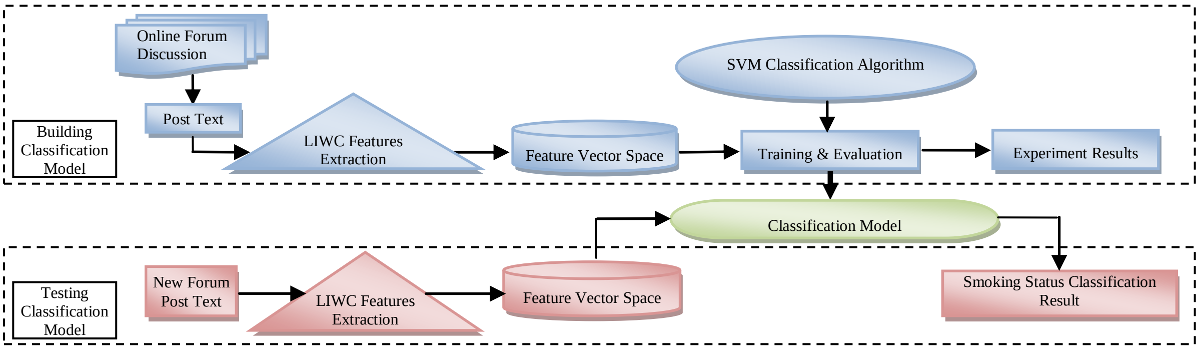


Figure 2: Framework of building and classifying UGC data with LIWC features set.

7000 posts, from publically available forums. To obtain enough covering data, a set of criteria was systematically applied on the collected data e.g. the availability of posts for those in-journey to stop smoking and frequency of posting. Furthermore, the same criteria used in identifying each class in the clinical reports corpus has been used in selecting the forum corpus, especially for the current smoker, past smoker and non-smoker classes. Different smoking words were used to retrieve online forum discussions such as; "smoker forum", "stop smoking forum".

In order to post to the online forums that had been selected, users must register. They have the option to enter a profile with pre-set fields. Unfortunately, not all profiles included a smoking status field. Also, not all users provided and updated their current smoking status. Furthermore, the smoking status extraction task is challenging in this research as no smoking keywords were relied on in classification. For these reasons, forum's title was dependent on annotating each post separately to be for past smoker, current smoker, non-smoker or people in-journey to stop smoking. The content of the posts was tokenised, converted to lower case, cleaned for non-English, repeated and empty posts. Online forum discussion corpus is freely available on http://dcapi.blogs.lincoln.ac.uk/?p=341.

Two different datasets were generated from forum discussions corpus for experimentation purposes, these were:

- “Dataset A" includes the same number of posts as the number of documents in the clinical reports corpus with maintaining the balance of the same number of documents in each smoking class similar to the clinical reports corpus;
- “Dataset B” was used in evaluating the effect of post length on the framework classification output. Depending on statistical calculations in forum corpus, in terms of the average of forum post length (word number), all posts that have less than 40 words were filtered out from the corpus. From the remaining forum posts, dataset B was generated with same properties as dataset A.

A clinical report dataset was requested directly from i2b2. It included four smoking classes (past smoker, current smoker, smoker and non-smoker) after filtering unknown classes from the corpus. After being tokenised and converted to lower case these reports were used in the experiments.

6.3 Results and Discussion

In this sub-section, we explain the experiments and present the results and discussion. This sub-section contains three experiments. The first is comparing the performance of forum and clinical report data (in baseline method), in terms of using binary word feature. The second experiment tests the suitability of the proposed feature type for classifying forum and clinical data. The third experiment is designed for examining the effect of post length, other classifier algorithm and filtering features on smoking status classification.

6.3.1 Compare with Baseline Method

The first experiment was performed to evaluate the performance of online forum discussion against the clinical reports (baseline method (Clark et al., 2008)) in the smoking status classification problem. The experiment used the baseline method, which achieved the best results in the smoking status challenge (Uzuner et al., 2008). The baseline method collected only binary unigram and bigram features and applied SVM classifier with 10-cross validation procedure to classify clinical reports, described in detail in Section 2. Forum dataset A, as described earlier, was used in this experiment.

Figure 3 shows the accuracy for the clinical reports and forum data over smoking status prediction. The figure shows that the binary word features are effective in classifying clinical records (82%) than forum data (75.69%). In addition, with binary word features type the size of the features vector space varies with the dataset size. For example, the binary feature vector length could be more than 20k, out of 390k words (5.128%) in a clinical report dataset, unless appropriate thresholding is applied. On the other hand, the binary feature vector length for the forum dataset can be more than 3K, out of 21K words (14.285%). When forum feature set was reduced to 1500, by using systematic method, the accuracy decreased to 60%. Furthermore, if new testing data/posts include new words that are unknown to the trained model, it will not be recogniszed and will not contribute to the classification resulting in potential reduction of accuracy.

Therefore, the use of the psycholinguistic features set (LIWC) is proposed to classify forum data, given its observed properties. This includes various categories (Section 4) and captures feelings, thoughts, emotions and experiences that arise in daily discussion for smoking status classification.

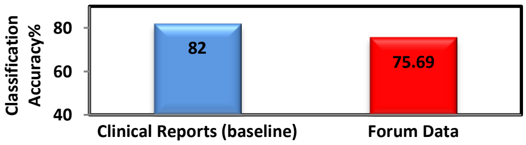


Figure 3: The classification accuracy results of classifying forum dataset A and clinical reports, when binary word features was used.

6.3.2 Feature Comparison

In order to analyse the selected feature set in classifying the proposed data, the framework in Figure 2 was executed in terms of using LIWC features set for classifying forum dataset A. The result was then compared with applying LIWC features on clinical reports data set. A feature vector space that represents the LIWC categories and subcategories values of each text in the forum and clinical reports datasets was built separately.

As shown in Figure 4, the forum data gives the highest accuracy of up to 75.09% whilst classifying with clinical report has a lower accuracy of 54%. Moreover, the feature vector length is both compact (88 features) and independent from the dataset itself. It mainly depends on the LIWC categories. This would also explain the results achieved. For example, with closer analysis we found that 70.99% of forum's words were identified by the LIWC dictionary, while only 38.93% words were recognized by the LIWC dictionary in the clinical report dataset. This can be attributed to the fact that online forum discussion is a wider discussion area than clinical reports and includes writing about feeling and emotions with discussions in different subjects which could cover different categories in LIWC dictionary.

Another observation is that using the LIWC feature for forum data has achieved slightly less accuracy compared to using binary word features on the same data. Nevertheless, the proposed features set is highly compact (only 88 features), independent of the dataset and fixed in length. Moreover, the LIWC contains psycholinguistic categories and covers different topics that could capture more words in forum discussions.

Further analysis has been done to extract significant LIWC categories that have been more positively affected during the classification process. Using the Principle Component Analysis (PCA) technique, main LIWC features were extracted from the forum's feature vector, as illustrated in Table 1. These features could be utilized for additional classification or clustering processes.

Another type of feature has been examined for smoking status classification (POS taggers). It mainly consists of 37 tags (The Stanford natural language processing group. October 2013). POS taggers was merged with LIWC features, thus the overall size of the feature set is 125 categories. Based on this combined new features set the accuracy increased slightly by 0.36%. Thus due to this small accuracy increase, we utilized LIWC features only for the forthcoming experiments, as it is also shorter (88 features). In addition, it is used for the first time for smoking status classification.

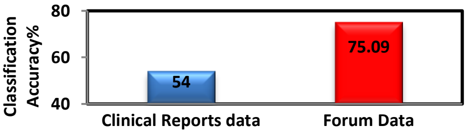


Figure 4: The classification accuracy results of classifying forum dataset A and clinical reports sets, when a LIWC features was used.

Table 1: Significant LIWC categories in forum's feature vector that impact positively on classification results, ranked according to PCA.

Discrepancy	Causation	Insight	Certainty	Sensory process
Inhabitation	Tentative	Optimism	Negative emotion	Positive emotion

In general, the new proposed feature could form the basis for further intensive investigation of a person's emotion and psychological state at various stages of the stop smoking process (or a similar task).

6.3.3 Factors Affecting the Results

The purpose of this subsection is to study the effect of different factors on the smoking status classification performance. These factors include the post length, applying different classifier algorithms and filtering features. These experiments used the framework explained in Figure 2, and datasets A and B, as described earlier.

6.3.3.1 Post Length

The effect of post length on smoking status classification was tested. Figure 5 represents the accuracy of framework when forum dataset B (includes posts with 40 words or more) was used as input data against result of classifying forum dataset A. It shows that when the post's length was increased the classification accuracy improved to 78.99%. This is due to the selected posts including general discussion and not specifically discussions related to smoking only. Thus, classifying posts that have more words gives more opportunity to extract valuable information that could help in inferring the smoking status. For example, in this experiment, forum dataset B with longer post scores a higher percentage of words that are identified by LIWC dictionary (72%) against forum dataset A before filtering shorter posts (70.99%).

In the selected forum corpus, classifying longer posts (in number of words) reflected higher accuracy. However, this is not a generalization, as the extra words should be relevant and recognizable by the utilized dictionary.

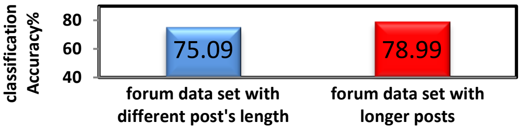


Figure 5: The classification accuracy results of classifying forum dataset A and B.

6.3.3.2 Different Classifiers

Different classifier algorithms were evaluated to predict the smoking status for the forum dataset A. KNN (K-nearest neighbours algorithm) and NB classifiers algorithms were selected (they were used before in literature for similar task) and trained on a LIWC features vector. The results of using other classifiers were compared with SVM output in Figure 6. They show SVM classifier giving higher accuracy against others. This result verifies (Sordo and Zeng, 2005; Wu et al., 2008) conclusion, that SVM algorithm is considered as one of the best machine learning algorithms in data mining. In addition, it is more accurate classifier algorithm to use for smoking status classification in forum data than KNN and NB.

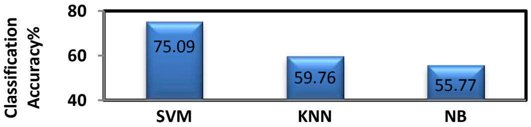


Figure 6: The classification accuracy results of classifying forum dataset A with different classifier algorithms.

6.3.3.3 Filtering Features

To examine the proposed feature set, this method was designed by removing redundant features that carry little information and do not assist in smoking status classification. Manual and systematic methods were followed to remove part of the LIWC features from the forum dataset A feature vector.

Manual feature selection method was performed by progressively removing weakest features by using Weka. The final selected group of features was 77 categories, whilst retaining the significant LIWC categories set that were extracted in (Section 6.3.2, Table 1). In the systematic method, the relation between each feature in the feature vector space and the smoking status classes was found. Potentially noisy features were then removed by using Weka’s chi-squared correlation weight. Figure 7 shows the effect on accuracy after removing features (by using both ways separately) and using a SVM classifier algorithm with the 10 cross-validation techniques.

However, contrary to expectations, these methods did not find a significant accuracy improvement against their original values. Filtering features resulted inreductions of their original value in about 0.19% and 2.39% in manual and systematic method respectively. This is because online forums include general discussions that involve most LIWC categories. Therefore when we removed part of these features from the LIWC feature set, the percentage of words that was recognized by LIWC dictionary decreases of 1.69% and 1.73% of their original values in manual and systematic methods respectively. In the systematic selection, the accuracy result was less than manual selection because there is more opportunity to remove essential categories (Table 1) from LIWC features list.

This finding suggests that the full set of the proposed features (88 categories and subcategories) is important for understanding and classifying online forum discussion.

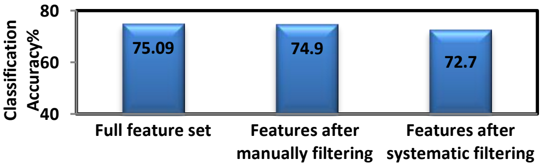


Figure 7: The classification accuracy results of classifying forum dataset A before and after filtering LIWC features set.

7 CONCLUSION AND FUTURE WORK

In this paper an investigation into the feasibility of text analysis has been presented to detect the smoking status in online forum discussions. Based on the investigation performed in this study, the contribution can be viewed from three aspects. The first contribution is analyzing online forum discussions that are widely available and easier to obtain, compared to clinical records. The second contribution is the utilization of psycholinguistic (LIWC) features. To the best of our knowledge, this is the first time that this feature set has been used for smoking status classification. Finally, the effects of post length, classifier algorithm and filtering features on the framework of smoking status classification were analyzed.

The experiments and results showed that, using the proposed features, the classification accuracy on the forum posts outperformed those on the clinical reports with LIWC features. Secondly, the proposed LIWC feature set has a fixed length, compact size (only 88 features), and is independent of the dataset size.

Thirdly, the result also established that an increase in post length (or the number of words in post) contributes to improving the classification accuracy. Although good and promising results were achieved using the proposed data type (online forum discussion) and features (LIWC) many directions remain open for development in this area of research. One of the important aspects is the utilisation of the LIWC dictionary for further analysis of a person's emotional and psychological status at various stages of the stop smoking process (i.e. in journey to stop smoking).

REFERENCES

Linguistic inquiry and word count. October 2013 Available from http://www.liwc.net/.

The Stanford natural language processing group. October 2013 Available from http://nlp.stanford.edu/software/tagger.shtml.

WEKA, the university of Wekato. October 2013 Available from http://www.cs.waikato.ac.nz/ml/weka/.

Informatics for integrating biology and the bedside. October 2013Available from https://www.i2b2.org/.

Aramaki, Eiji, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. 2006. Patient status classification by using rule based sentence extraction and BM25 kNN-based classifier. Paper presented at i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, [no pagination] .

Clark, C., K. Good, L. Jezierny, M. Macpherson, B. Wilson, and U. Chajewska. 2008. Identifying smokers with a medical extraction system. *Journal of the American Medical Informatics Association* 15 (1): 36-39.

Cohen, Aaron M. 2008. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *Journal of American Medical Informatics Association* 15 (1): 32-35.

Gill, Alastair J., Scott Nowson, and Jon Oberlander.2006.Language and personality in computer-mediated communication: A cross-genre comparison. *Journal of Computer Mediated Communication*, [no pagination].

Kaiser, C., and F. Bodendorf. 2012. Mining patient experiences on web 2.0-A case study in the pharmaceutical industry. Paper presented at SRII Global Conference (SRII), 2012 Annual, 139-145 .

Leshed, Gilly, and Joseph'Jofish' Kaye. 2006. Understanding how bloggers feel: Recognizing affect in blog posts. Paper presented at CHI'06 extended abstracts on Human factors in computing systems, 1019-1024.

Liu, Mei, Anushi Shah, Min Jiang, Neeraja B. Peterson, Qi Dai, Melinda C. Aldrich, Qingxia Chen, Erica A. Bowton, Hongfang Liu, and Joshua C. Denny. 2012. A study of transportability of an existing smoking status detection module across institutions. Paper presented at AMIA Annual Symposium Proceedings, 577-586.

Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. Paper presented at Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 79-86.

Pedersen, Ted. 2006. Determining smoker status using supervised and unsupervised learning with lexical features. Paper presented at i2b2Workshop on Challenges in Natural Language Processing for Clinical Data, [no pagination].

Savova, Guergana K., Philip V. Ogren, Patrick H. Duffy, James D. Buntrock, and Christopher G. Chute. 2008. Mayo clinic NLP system for patient smoking status identification. *Journal of American Medical Informatics Association* 15 (1): 25-28.

Sordo, Margarita, and Qing Zeng. 2005. On sample size and classification accuracy: A performance comparison. In *Biological and medical data analysis*., 193-201Springer.

Szarvas, György, RichárdFarkas, SzilárdIván, AndrásKocsor, and RóbertBusaFekete. 2006. Automatic extraction of semantic content from medical discharge records. Paper presented at i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, [no pagination].

Tausczik, Y. R., and J. W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29 (1): 24-54.

Uzuner, Özlem, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal American Medical Informatics Association* 15 (1): 14-24.

Wicentowski, Richard, and Matthew R. Sydes. 2008. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. *Journal of the American Medical Informatics Association* 15 (1): 29-31.

Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. Yu. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14 (1): 1-37.

Zeng, Q. T., S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making* 6 (1): 30-38.