A preliminary study on automatic identification of patient smoking status in unstructured electronic health records

Jitendra Jonnagaddala
School of Public Health and Community Medicine, University of New South Wales, Australia
z3339253@unsw.edu.au

Hong-Jie Dai*
Department of Computer Science and Information Engineering, National Taitung University, Taiwan
hjdai@nttu.edu.tw

Pradeep Ray
Asia-Pacific Ubiquitous Healthcare Research Centre, University of New South Wales, Australia
p.ray@unsw.edu.au

Siaw-Teng Liaw*
School of Public Health and Community Medicine , University of New South Wales, Australia
siaw@unsw.edu.au

Abstract

Identifying smoking status of patients is vital for assessing their risk for a disease. With the rapid adoption of electronic health records (EHRs), patient information is scattered across various systems in the form of structured and unstructured data. In this study, we aimed to develop a hybrid system using rule-based, unsupervised and supervised machine learning techniques to automatically identify the smoking status of patients in unstructured EHRs. In addition to traditional features, we used per-document topic model distribution weights as features in our system. We also discuss the performance of our hybrid system using different feature sets. Our preliminary results demonstrated that combining per-document topic model distribution weights with traditional features improve the overall performance of the system.

1 Introduction

Electronic health records (EHRs) carry vital patient information. EHRs generally store information such as medical history, procedures and tests, medications, admissions data and social history. Social history includes details on a patient's smoking habits, alcohol and drug usage. However, most of the information stored in EHRs are in the free-text form as clinical narratives. Natural language processing (NLP) and text mining can be used to extract this valuable information from unstructured EHRs. The extracted information in turn can be used to build a number of applications such as clinical decision support, medical coding, cohort selection and registry systems (Jensen, Jensen, & Brunak, 2012; Jonnagaddala, Dai, Ray, & Liaw, 2015).

Smoking is known to be one of the major risk factors in the development of coronary artery disease, cardiovascular disease, chronic kidney disease and cancer. Thus, identifying smoking status automatically from unstructured EHRs is crucial for preventive medicine. Smoking status can be used to assess risk for a particular disease and provide interventions based on clinical guidelines (Jonnagaddala, Liaw , et al., 2015). Identifying smoking status automatically in unstructured EHRs is not straightforward and often complex. Clinicians usually report smoking information in various formats. For example, few clinicians report in packs per day and others simply classify patient as just smoker or non-smoker.

Previous studies have reported success in using support vector machines (SVMs) to automatically identify smoking status in unstructured EHRs (Clark et al., 2008; Cohen, 2008; Khor et al., 2013; Savova et al., 2010; Savova, Ogren, Duffy,

Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 147–151,
Beijing, China, July 30, 2015.

Buntrock, & Chute, 2008). Similarly, Bui et al developed a system using SVMs by automatically learning regular expressions from two different datasets (Bui & Zeng-Treitler, 2014). However, most of these studies developed their automated systems using traditional features like unigrams, bigrams and POS tags in combination with few rules (Uzuner, Goldstein, Luo, & Kohane, 2008). In this study, we developed a hybrid system using topic modelling and SVMs to automatically identify patients smoking status in unstructured EHRs. Per-document topic distribution weights obtained from unsupervised topic modelling technique are used as features together with traditional features. For the purpose of this study we combined two different datasets to form one large dataset. The system classifies patients into five categories depending on their smoking history using rule-based and machine learning techniques.

2 Materials and Methods

2.1 Dataset

The dataset used in the study is generated by merging datasets from the 2006 and 2014 NLP challenges set forth by the information for integrating biology to the bedside (i2b2) project(Amber Stubbs, Kotfila, Xu, & Uzuner, 2015; A. Stubbs & Uzuner, 2015; Uzuner et al., 2008). The 2006 i2b2 dataset has one document per patient. The 2014 has multiple documents (from multiple encounters) per patient. In this study, we aim to identify the smoking status of a given document irrespective of the fact that one patient might have multiple documents with varying smoking status. In other words, we aimed to develop an automated system to identify smoking status at document level. The final merged dataset consisted of documents classified into one of the five possible smoking categories listed below:

Current Smoker: A current smoker class is assigned to a document when it explicitly state that the patient was a smoker within the past year. If the document mentions, patient has quit smoking within the past one year, the document is still classified as current smoker.

Past Smoker: A past smoker is when a document explicitly state that the patient used to smoke more than a year ago.

Past or Current Smoker: A past or current smoker is assigned when a document mentions that patient smokes, but not possible to determine the status either as past or current.

Non-Smoker: A non-smoker is when documents explicitly states that they never smoked.

Unknown: An unknown status is assigned to a document if there is no mention of smoking.

2.2 Baseline System

The smoking status classifier of nttmuClinical.NET (Chang, Dai, Jonnagaddala, Chen, & Hsu, 2015) was used as the baseline system in this study. For the detection of smoking status, a list of smoking-related keywords, such as "smoking" and "cigarette", was matched with the given document by the classifier. If no match was found, the document was automatically assigned with the UNKNOWN class. Otherwise, the line containing the listed terms was regarded as a

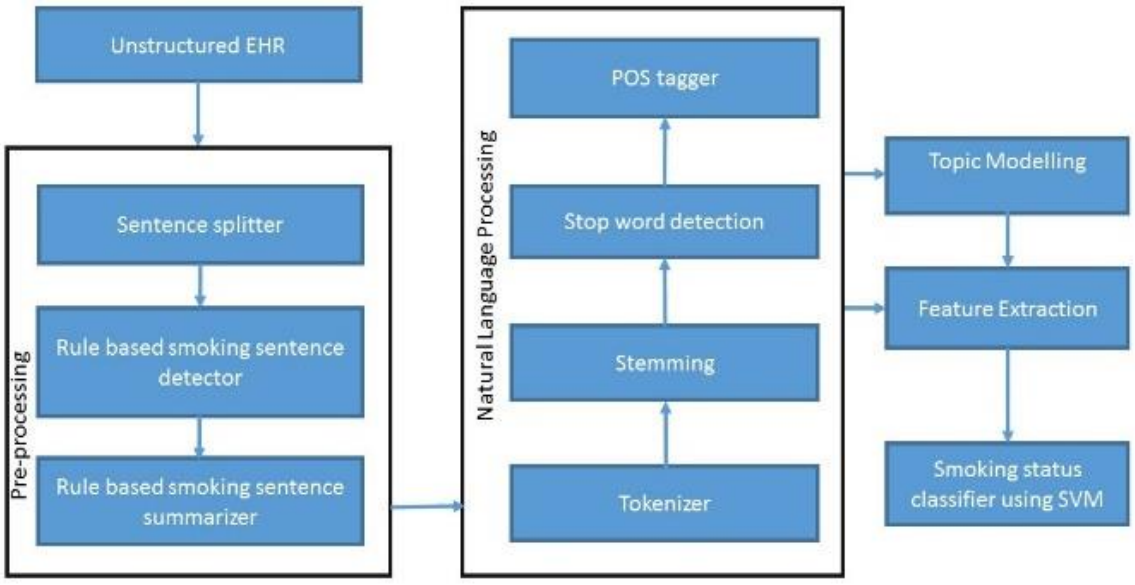

Figure 1: Overview of components in smoking identification pipeline

context that could provide more information for detecting the smoking status of a patient in the document. The context-aware algorithm with several weighted rules developed by leveraging document creation information for different smoking statuses was then applied on the document to determine the smoking status. The algorithm starts by checking the current context. If the context did not provide sufficient temporal information to determine the smoking status, the classifier extends the current context to include more sentences and re-apply the developed rules until either the status was determined or no further updated context was available.

2.3 Hybrid system for smoking classification

Our smoking status identification system takes advantage of the fact that, most of the documents had smoking related information present in a particular section of the document. Thus, instead of using the whole document for classification, we first extracted the smoking related sentences and then used those sentences to identify smoking status. Our system comprised of the following components (Figure1).

Sentence Splitter: To split the documents into individual sentences we used sentence segmentation available in Stanford coreNLP (Manning et al., 2014). The tool was modified to preserve the section headers like "Family History" and "Social History".

Smoking sentence detector: This component was developed to extract the smoking related sentences from the documents. The component identified the smoking status related terms and extracted surrounding sentences.

Smoking sentence summarizer: As some of the documents had multiple smoking related instances a rule-based component was developed to summarize these sentences. The rules were created based on the headers, like Social History, Habits etc.

NLP Component: Once the smoking related sentences were identified and summarized, they were processed further using multiple core NLP components - tokenizer, stemming, stop words removal and POS tagging to generate features.

Feature Extraction: After NLP was done, multiple feature sets were developed including unigrams, bigrams, POS bigrams, word POS pairs and topic models. We generated ten topics using Latent dirichlet allocation (LDA) and Gibbs sampling (Blei, Ng, & Jordan, 2003). The per-document distribution weights of the topics were later incorporated into the feature sets used to train smoking status classifier.

SVM Classifier: Linear SVM classifier was used to classify the documents into one of the five classes discussed above. The cost parameter was optimized to 0.01 for better performance. The SVM classifier was developed using training set and evaluated on test set. The performance of the developed system is presented in the form of precision (P), recall (R) and F1 score (F1) in micro and macro averaged settings.

3 Results

We observed that the 2006 and 2014 i2b2 NLP smoking datasets are not identical in structure and smoking classification classes. We implemented few changes to standardize the smoking status in the merged dataset. Similarly, we also manually annotated documents where smoking status was missing, even though available in documents. Where the smoking status cannot be determined we labeled them as unknown. The summary of number of documents available in final merged dataset (training and test) with the class distribution is presented in Table 1.

Smoking classification classes	Training	Test
Current Smoker	100	46
Past Smoker	185	124
Non-Smoker	251	136
Past Or Current Smoker	29	6
Unknown	623	306
Total no. of documents	1188	618

Table 1 Document level class distribution of dataset

The training set was processed through our hybrid system to generate features and train linear SVM classifier to perform multi class classification. Initially the training set generated model was evaluated using tenfold cross validation on same. This evaluation allowed us to tweak the parameters of our components for better performance. We also used grid search to identify best parameters for linear SVM. The results on the test set with best performing parameters are reported in Table 2. The feature set which incorporated topic modelling based features performed better than baseline and traditional feature set. The topic modelling based feature set trained SVM classifier achieved F1 measure of 83.66% whereas the traditional feature set achieved F1 measure of 82.69% and baseline system 81.85%.

Feature set	Micro averaged		
	P	R	F1
Baseline	0.8185	0.8185	0.8185
Unigrams, Bigrams, POS bigrams, Word POS pairs	0.8269	0.8269	0.8269
Unigrams, Bigrams, POS bigrams, Word POS pairs, Topic models	0.8366	0.8366	0.8366

Table 2: Micro averaged results on test set

4 Discussion

Linear SVMs were used in this study and during the development stage it was observed that the linear kernel performs better than non-linear kernels like radial basis function (RBF). The reason behind the better performance of the linear kernel may be attributed to the presence of a large number of features. It is also believed that when the number of features is much greater than the number of instances then mapping the feature space to a higher dimension like in RBF adds no improvement to the performance of the system.
We also noticed that adding topic models as features did increase the performance of the classifier. However, we believe that the overall performance of classifier can be further increased by optimizing the number of topics to be extracted. The high number of topics we chose to extract using LDA algorithm in current setting are creating sparse features for SVM classifier. Further investigation into choosing optimal number of topics is required.

Both training and test sets in the merged dataset included almost half of documents with unknown class. SVMs in general tend to be biased towards majority classes giving less priority to minority classes. This resulted in significant gap between micro and macro averaged scores. This problem can be solved by taking a multi layered classification approach. As the system is detecting smoking related sentences first, one of the ways to classify is to mark all the instances with no smoking reference as unknown and then classify the remaining into two groups smoker and non-smoker followed by past and current smoker. Another option to address this imbalance problem is by assigning weights to the SVM classifier (Chew, Bogner, & Lim, 2001). Our system also failed to classify current smoker and past smoker efficiently mainly due to negation. The performance of our system can be further improved by implementing a negation component in conjunction with temporal component which can leverage discharge/admission dates and document generated dates as demonstrated in the baseline system. During our error analysis we also noticed that few documents included smoking related administration data in the form of billing and medication codes. We can also use this information to improve the performance of our system (Wiley, Shah, Xu, & Bush, 2013).

5 Conclusion

In summary, we presented the results of a preliminary study in automatically identifying smoking status in unstructured EHRs using SVMs and topic models. Our approach encompassed usage of per-document topic distribution weights generated from topic modelling as features in conjunction with several other traditional features extracted from NLP pipeline. We compared the results of our system using various feature sets against a baseline system. The results demonstrated that topic modelling is useful in identifying smoking status, however, proper topic sampling strategies should be employed. Also, the need for the inclusion of negation and temporal information recognition components in smoking identification is highlighted. In future, we would like to improve our system performance by employing negation and temporal related features. We also would like to explore optimal topic size for smoking identification from relevant smoking related sentences and compare the performance of our system against various smoking identification systems available like Apache cTAKES (Savova et al., 2010).

Acknowledgements

The authors would like to thank the organizers of 2014 and 2006 i2b2/UTHealth Shared-Tasks. De-identified health records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by grants 2U54LM008748 and 1R13LM01141101 from National Institute of health (NIH). The authors would like to thank the organizers of 2014 i2b2/UTHealth Shared-Tasks. De-identified health records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by grants 2U54LM008748 and 1R13LM01141101 from National Institute of health (NIH). This study was conducted as part of

the electronic Practice Based Research Network (ePBRN) and Translational Cancer research network (TCRN) research programs. ePBRN was/is funded in part by the School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research, UNSW Medicine and South West Sydney Local Health District. TCRN is funded by Cancer Institute of New South Wales and Prince of Wales Clinical School, UNSW Medicine. The content of this publication is solely the responsibility of the authors and does not necessarily reflect the official views of the funding bodies.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Bui, D. D. A., & Zeng-Treitler, Q. (2014). Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association, 21*(5), 850-857.

Chang, N.-W., Dai, H.-J., Jonnagaddala, J., Chen, C.-W., & Hsu, W.-L. (2015). A Context-Aware Approach for Progression Tracking of Medical Concepts in Electronic Medical Records. Manuscript submitted.

Chew, H.-G., Bogner, R. E., & Lim, C.-C. (2001). *Dual v-support vector machine with error rate and training size biasing.* Paper presented at the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01).

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

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

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics, 13*(6), 395-405.

Jonnagaddala, J., Dai, H.-J., Ray, P., & Liaw, S.-T. (in press). Mining electronic health records to guide and support good clinical decision support systems. In J. Moon & M. P. Galea (Eds.), *Improving Health Management through Clinical Decision Support Systems*: IGI-Global.

Jonnagaddala, J., Liaw , S.-T., Ray , P., Kumar , M., Chang, N.-W., & Dai, H.-J. (2015). Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of Biomedical Informatics*.

Khor, R., Yip, W.-K., Bressel, M., Rose, W., Duchesne, G., & Foroudi, F. (2013). Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. *Journal of the American Medical Informatics Association*, amiajnl-2013-002090.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). *The Stanford CoreNLP natural language processing toolkit.* Paper presented at the Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association, 17*(5), 507-513.

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., & Chute, C. G. (2008). Mayo clinic NLP system for patient smoking status identification. *Journal of the American Medical Informatics Association, 15*(1), 25-28.

Stubbs, A., Kotfila, C., Xu, H., & Uzuner, O. (2015). Practical applications for NLP in Clinical Research: the 2014 i2b2/UTHealth shared tasks.

Stubbs, A., & Uzuner, O. (2015). Annotating Risk Factors for Heart Disease in Clinical Narratives for Diabetic Patients. *J Biomed Inform*. doi: 10.1016/j.jbi.2015.05.009

Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association, 15*(1), 14-24.

Wiley, L. K., Shah, A., Xu, H., & Bush, W. S. (2013). ICD-9 tobacco use codes are effective identifiers of smoking status. *Journal of the American Medical Informatics Association, 20*(4), 652-658.