Classification of Interviews – A Case Study on Cancer Patients

Braja Gopal Patra[1] *Amitava Kundu*[1] *Dipankar Das*[2] *Sivaji Bandyopadhyay*[1]
(1) JADAVPUR UNIVERSITY, Kolkata, India
(2) NATIONAL INSTITUTE OF TECHNOLOGY, Meghalaya, India
brajagopal.cse@gmail.com, amitava.jucse@gmail.com,
dipankar.dipnil2005@gmail.com, sivaji_cse_ju@yahoo.com

ABSTRACT

With the rapid expansion of Web 2.0, a variety of documents abound online. Thus, it is important to find methods that can annotate and organize documents in meaningful ways to expedite the search process. A considerable amount of research on document classification has been conducted. However, this paper introduces the classification of interviews of cancer patients into several cancer diseases based on the features collected from the corpus. We have developed a corpus of 727 interviews collected from a web archive of medical articles. The TF-IDF features of unigram, bigram, trigram and emotion words as well as the SentiWordNet and Cosine similarity features have been used in training and testing of the classification systems. We have employed three different classifiers like *k*-NN, Decision Tree and Naïve Bayes for classifying the documents into different classes of cancer. The experimental results obtain maximum accuracy of 99.31% tested on 73 documents of the test data.

KEYWORDS: TF-IDF, document classification, cancer patients, emotion words and SentiWordNet.

Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), pages 27–36,
COLING 2012, Mumbai, December 2012.

1 Introduction

With the explosion of online electronic documents in recent times, document classification is becoming the necessary assistance to people in searching, organizing and collecting related documents. The task of automatic classification is a classic example of pattern recognition, where a classifier assigns labels to the test data based on the labels of the training data. Document classification is the task of assigning a document to one or more classes.

The present paper reports a task of classifying interviews of cancer patients. The primary objective is to predict the type of cancer, given a particular interview of a patient. We have developed a corpus from an open source web archive [1] of interviews conducted only for the cancer patients. The interview documents of the corpus are stored in XML format and pertaining to a total number of 17 classes of cancer. Three classifiers, namely *k*-NN, Naïve Bayes and Decision Tree were used for document classification based on TF-IDF scores and Cosine similarity. We have calculated the TF-IDF scores of unigrams, bigrams, trigrams, and emotion words. As a part of the experiment, the clustering was done by considering the similar hypernym sets of two words. We have used the scores of the word groups or WordNet clusters instead of using individual words only.

A considerable amount of research on document classification has already been conducted by different research groups such as the machine learning techniques (Sebastiani, 2002) have been adopted with great effect whereas Li and Jain, (1998) provides a brief overview of document classification. It has been observed that the Decision tree classifiers, nearest neighbor algorithms, Bayesian classifiers and support vector machines have been common choice of researchers and have produced satisfactory results. The idea is to extract the features from each document and then feed them to a machine learning algorithm (Dumais et al., 1998; Joachims, 1998). The Bag of words features such as document frequency (df) and TF-IDF features (Han et al., 2000) yield decent accuracies. Yang and Wen, (2007) achieved a maximum accuracy of 98% using TF-IDF scores of bag of words.

Enabling convenient access to scientific documents becomes difficult given the constant increase in the number of incoming documents and extensive manual labor associated with their storage, description and classification. Intelligent search capabilities are desirable so that the users may find the required information conveniently (Rak et al., 2005). This is particularly relevant for repositories of scientific medical articles due to their extensive use, large size and number and well maintained structure. The authors also report an associative classification of medical documents. Uramoto et al., (2004) developed MedTAKMI (Text Analysis and Knowledge Mining for Biomedical Documents), an application tool to facilitate knowledge discovery from very large text databases. However, the present task focuses on interview articles of the patients in cancerous conditions, as an initial step of developing a larger corpus of medical articles. It also attempts to discover semantically related word clusters from the collected interview documents. Additionally, it reports the affective statistics of the corpus and attempts to relate the frequency of emotional responses to the type of ailment of the patient. The overall aim of this research is to identify the textual clues which help in diagnosing the symptoms of the patients from their interviews. We have also incorporated the sentiment related hints so as to boost the identification of symptoms with focused perspectives.

[1] http://www.healthtalkonline.org/

The rest of the paper is organized in the following manner. Section 2 provides details of resource preparation. Section 3 provides an elaborative description of the features used in the task. Next, Section 4 describes the implementation of machine learning algorithms while Section 5 presents the results and analysis. Finally, conclusions and future directions are presented.

2 Resource preparation

Healthtalkonline is an award winning website that shares more than 2000 patients' experiences of over 60 health-related conditions and ailments. For our present task, we have prepared a corpus of 727 interviews of cancer patients collected from the above mentioned website. We have developed a web crawler which has been used to collect the data available on the www.healthtalkonline.org website. Once the URL of an arbitrary webpage containing a cancer related interview was supplied to the crawler, it was able to hop all other pages containing the cancer-related interviews. As such, URLs of all the webpages containing cancer interviews were spotted and thereafter, data was extracted from these pages. An initial manual examination revealed that the webpages have different formats. Thus, three kinds of patterns were observed. All unnecessary information was eliminated and the refined data was stored in XML format. A snapshot of a portion of such a XML document is shown in Figure. 1. The statistics of the corpus are given in Table 1. Out of the 727 XML documents prepared, 85% were used as training data, 5% for development and the rest 10% were used as test data. The corpus contains interviews only and is thus comprised of questions and the corresponding answers. Each line of an actual interview is either a narration/question indicative of the patient's conditions or is a response from the patient.

```
<Body>
<Situation id="1">
Aley was diagnosed with biphenotypic acute leukaemia (BAL), a mixture of myeloid and lymphoblastic leukaem
</Situation>
<QuestionAns id="1">
The day I received a call from the blood transfusion centre and a doctor said to me that I had to go and s
</QuestionAns>
<Situation id="2">
Aley broke the news to his brother in Pakistan in stages so as to prepare the family for the bombshell of
</Situation>
<QuestionAns id="2">
Then the following day I called my brother. He is older than me, three years older than me but we have got
</QuestionAns>
<Situation id="3">
Aley's treatment would definitely cause infertility but he was more concerned about curing the leukaemia t
</Situation>
<QuestionAns id="3">
Yes they did tell me and they have to store your sperm for if, because I do not have any children. I was n
</QuestionAns>
<Situation id="4">
Aley will have his white blood cells extracted and treated with UV light to counter the rash he has develo
</Situation>
```

FIGURE 1 – A Snapshot of an interview document in XML format.

Prior to feature extraction, the corpus was tokenized. Separate lists of unigrams, bigrams, trigrams, emotion words were prepared. Emotion words were identified using a SentiWordNet[2] lexicon. Some words are abundant and have little semantic content known as stop words. There were 329 stop words prepared by us manually. They were removed from the list of tokens actually utilized. Also, named entities were assumed to have little role to play in classification

[2] http://sentiwordnet.isti.cnr.it/

and hence were excluded too. Named entities were identified using the Stanford Named Entity Recognizer version 1.2.6[3]. Lists of semantically related clusters of words were prepared using hypernyms of each unigram.

Total number of words after removing stop words	421867
Total number of unique words	17627
Total number of named entity	182
Total number of Emotion words identified using SentiWordNet lexicon	5900
Total number of emotion words occurred more than three documents	3091
Total number of word classes after clubbing similar words	11466
Total number of word classes after clubbing similar words more than four documents	6287
Total number of Bigrams	201729
Total number of Bigrams occurred more than four documents	8286
Total number of Trigrams	285993
Total number of Trigrams occurred more than three documents	22082

TABLE 1 –Statistics of corpus.

3 Feature Selection

Feature selection plays an important role in machine learning framework and also in automatic document classification. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately whereas feature selection is the process of removing the irrelevant and redundant features and reducing the size of the feature set for better accuracies. The following experiments have been carried out to find out suitable features. We have used TF-IDF and Cosine Similarity feature vectors. TF-IDF of emotion words, unigrams, bigrams and trigrams have also been considered in feature vectors. The dimensionality of the feature vector space is very high. To reduce the dimensionality, semantically related unigrams were first clustered using hypernyms and then TF-IDF scores of these word groups were considered.

3.1 TF-IDF

TF-IDF is the most common weighting method which reflects the importance of each word in a document. It describes the document in a Vector space model and is used in Information Retrieval and Text Mining (Soucy and Mineau, 2005). A document is represented as the pair <t,w>, where t= {t_1, t_2, t_3… t_n} is the set of terms and w={w_1, w_2, w_3,…w_n} is the set of corresponding TF-IDF weights of the terms. The TF-IDF weight can be computed as follows

$$w_i = \begin{cases} \log TF(t_i,d) \times IDF(t_i) & if\ TF(t_i,d) \geq 1 \\ 0 & otherwise \end{cases}$$

Where TF(t_i, d) is the frequency of the term t_i in the document d.

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

$$IDF(t_i) = \log(\frac{|D|}{DF(t_i)})$$

|D| is the total number of documents and DF(t_i) is the number of documents in which the term t_i is present.

3.1.1 TF-IDF of emotion words (TF-IDF$_{emo}$)

In our corpus of cancer patients' interviews, a lot of emotional responses were observed. Each interview was replete with emotion words. We have identified the emotion words from the list of unigrams using the SentiWordNet lexicon and then computed TF-IDF of these emotion words as a feature set. The aim was to find any correlation of frequency of emotion words in an interview to the severity/kind of ailment. In fact, a relatively higher number of occurrences of emotion words were observed in interviews related to certain kinds of cancer. Figure 2 illustrates the same.

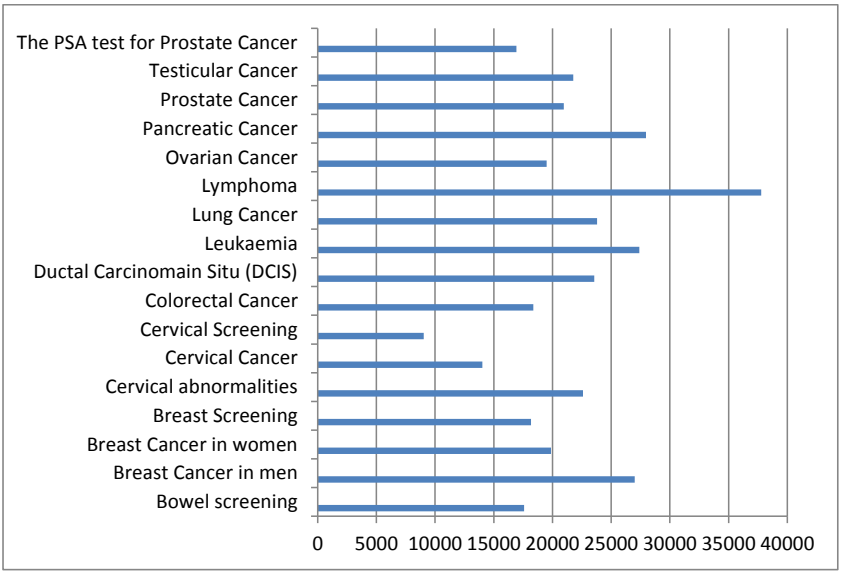

FIGURE 2 – Frequency of emotion words per cancer category

3.1.2 TF-IDF of frequent words using WordNet similarity (TF-IDF$_{wc}$)

A manual observation suggests that the words having document frequency more than four can be considered as the more useful features for our experiment. Hence, we have clubbed or clustered the similar words using hypernym relation. A hypernym is a word or phrase whose referents form a set including as a subset the referents of a subordinate term. In other words, hypernym refers to a broader meaning for a class of words. For example, 'colour' is a hypernym of 'red'. The feature vector is very large as it includes all types of features. So the feature vector is needed to be reduced to get a better result. Thus, by clustering the semantically related words, we can reduce

the size of the feature vector. For the purpose, hypernyms of every unstemmed unigram were obtained by using the RitaWordnet[4] library methods.

3.1.3 TF-IDF of bigrams ($TF\text{-}IDF_B$)

We have listed the bigrams of the entire corpus. It has been found that if a bigram occurs in at least five documents, it is an effective feature in classification of documents. We have found a total of 201729 numbers of bigrams and out of these 8286 numbers of bigrams occurred in five or more documents, as shown in Table 1.

3.1.4 TF-IDF of trigrams ($TF\text{-}IDF_T$)

We have also listed the trigrams of the entire corpus. Those trigrams that occur in more than three documents have been are identified as effective features. We found a total of 285993numbers of trigrams and out of these 22082 trigrams occurred in more than three documents, as shown in Table 1.

3.2 Cosine Similarity

Cosine similarity is one of the most commonly used metrics deployed to find out the similar documents (Han et al., 2001; Dehak, 2010) from a large pool of documents. Cosine similarity is particularly effective in finding out similar documents in a high dimensional feature space. The advantage of using Cosine Similarity is that it is normalized and lies in [0,1]. Given a vocabulary V, each document d can be represented as a vector as follows:

$$\mathbf{d} = <tf(t_1,d),tf\ (t_2,d),\ldots,tf(t_{|V|},d)>$$

where $t_1,t_2,..t_{|V|}$ are the words of vocabulary V.

Given two document vectors $\mathbf{d_i}$ and $\mathbf{d_j}$, the cosine similarity between them is computed as follows

$$\cos(d_i, d_j) = \frac{\sum_{w \in V} tf(w, d_i) * tf(w, d_j)}{\sqrt{\sum_{w \in V} tf(w, d_i)^2}\sqrt{\sum_{w \in V} tf(w, d_j)^2}}$$

4 Classification of Documents

We have used three types of classifiers namely *k*-NN, Naïve Bayes and Decision Tree. The dimensionality of the feature vector space is quite high while the number of documents available in the corpus is less. Therefore, we have carried out our experiments using the above novel classifiers only instead of more complicated ones like SVM or CRF. The system architecture is given below in Figure. 3, which illustrates the steps of the task.

The corpus was first developed and cleansed during pre-processing. Thereafter, various features were extracted and the resulting data was fed to machine learning algorithms. We have used Weka 3.7.7[5] for our classification experiments. Weka is an open source data mining tool. It presents collection of machine learning algorithms for data mining tasks. 85% of the corpus was used as training data, 5% as development set and the rest 10% as test data. In order to obtain reliable accuracy, a 10-fold cross validation was performed for each classifier.

[4]http://rednoise.org/rita/wordnet/documentation/index.htm
[5] http://www.cs.waikato.ac.nz/ml/weka/

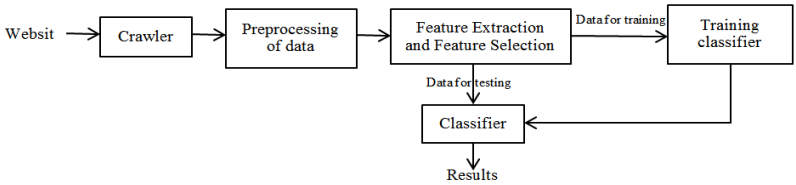

FIGURE 3 – System Architecture.

4.1 *k*-NN classifier

The documents have been represented as vectors in the TF-IDF vector space. The k-NN classifier is a learning algorithm which uses the labels of neighboring training points to decide the label of a test point. In fact, class labels of k-nearest neighbors are considered. We have used the IBk algorithm, which is an implementation of k-NN algorithm in Weka. Higher values of k are most likely to reduce the effect of outliers. However, we have used k=3 in our experiments.

4.2 Naïve Bayes classifier

Naïve Bayes is a simple probabilistic classifier based on the Bayes theorem and strong independence assumptions. The naïve Bayes model is tremendously appealing because of its robustness, elegance and simplicity. Despite being one of the oldest formal classification algorithms, it often is surprisingly effective even in its simplest forms. It is widely used for text classification purposes. We have used the multinomial Naïve Bayes classifier available in Weka tool. It is a specialized Naïve Bayes classifier for text classification.

4.3 Decision Tree classifier

We have used the J48 decision tree classifier available in the Weka tool for our purpose. J48 is actually a Java implementation of the C4.5 algorithm. C4.5 produces an initial decision tree and implements pruning in order to avoid over fitting.

5 Evaluation Results and Discussion

The interview corpus contains documents pertaining to a total of 17 classes of cancer disease. On an average, there are 40 documents per category of cancer. To obtain reliable results, we have performed a 10-fold cross validation for each of the classification experiments. The ablation study has been performed by including each of the features, separately for classification. Results obtained using bigrams, trigrams, word clusters and emotion words features have also been recorded in Table 2. The outcomes also reflect the combined effect of different features. Table 2 presents the accuracy, precision, recall and F-score of each classifier for different feature sets.

On the other hand, in a bag of words model, the Cosine similarity of documents has been considered as a feature using frequency of words only. It has been observed from the outcomes of the experiments that the Cosine similarity produces modest accuracies whereas the emotion word feature produces low accuracies compared to other TFIDF results. It is found that a total of 3091 emotion words are present on an average of three documents.

		Cosine Similarity	TF-IDF$_{emo}$	TF-IDF$_{wc}$	TF-IDF$_{B}$	TF-IDF$_{T}$	TF-IDF$_{emo}$ +TF-IDF$_{wc}$	TF-IDF$_{emo}$ +TF-IDF$_{wc}$ + TF-IDF$_{B}$	TF-IDF$_{emo}$ +TF-IDF$_{wc}$ +TF-IDF$_{B}$ +TF-IDF$_{T}$
Accuracy	*k*-NN	36.45	15.13	24.48	51.44	86.93	23.1	48.0	61.7
	Naïve Bayes	43.33	66.16	88.17	97.11	97.52	88.17	96.97	98.62
	Decision Tree	37.42	65.2	92.84	98.48	99.31	92.84	98.7	98.62
Precision	*k*-NN	38.2	28.8	53.2	67.1	91.4	47.4	56.5	77.6
	Naïve Bayes	53.5	67.5	88.9	97.2	97.6	89.0	97.1	98.7
	Decision Tree	37.4	65.9	93.0	98.5	99.3	93.0	98.7	98.6
Recall	*k*-NN	36.5	15.1	24.5	51.4	86.9	23.1	48.0	61.8
	Naïve Bayes	43.3	66.2	88.2	97.1	97.5	88.2	97.0	98.6
	Decision Tree	37.4	65.2	92.8	98.5	99.3	92.8	98.6	98.6
F-score	*k*-NN	35.6	10.7	21.9	46.0	87.5	20.2	42.1	56.3
	Naïve Bayes	40.7	66.0	88.2	97.1	97.5	88.2	97.0	98.6
	Decision Tree	37.3	65.3	92.8	98.5	99.3	92.8	98.6	98.6

TABLE 2 –Result of the experiments (in %).

It is observed that by considering the TF-IDF of word clusters, we achieved moderate accuracies and especially the bigram features produce satisfactory results. It has also been observed that the bigrams that occur in at least five documents are most informative and produce best results. The accuracies fall when bigrams occurring in more than five documents are considered, seemingly because of the reduced number of features. The trigram features have been found to be most informative feature and produce best accuracies. We have considered only those trigrams that occur in at least three documents. Another experiment has been conducted using the combined features of emotion words and word clusters. In this case, the accuracies produced are comparatively lower than that produced by trigram features. When bigram features are combined with former two features, accuracies have been improved. It has to be mentioned that further improvement in accuracies were also observed after adding the trigram features. The k-NN classifier produced low accuracies overall and J48 decision tree produces best accuracies overall.

Conclusion and future work

In this work, we have presented a task of classifying cancer patients' interviews into different types of cancer. We have performed our experiments on a corpus of 727 interview documents extracted from the healthtalkonline website. As a part of our future work, we intend to expand

our corpus and include articles related to all other ailments available on the website as well. The features in the experiments have produced decent results. Maximum accuracy of 99.31% has been obtained using trigram features. Having observed abundant occurrences of emotion words in our corpus, we are planning to use our corpus for further affective analysis. Thus, our aim is to extract the patient's responses separately. In the present work, a bag of words model has been used whereas the identification of more informative features and dimensionality reduction remains another objective.

Acknowledgments

The work reported in this paper is supported by a grant from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.

References

Danisman, T. and Alpkocak, A. (2008). Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, Vol. 2, pages 53-59.

Dehak, N., Dehak, R., Glass, J., Reynolds, D. and Kenny, P. (2010). Cosine similarity scoring without score normalization techniques. In *Proceedings of Odyssey Speaker and Language Recognition Workshop*.

Dumais, S., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148-155. ACM.

Han, E. H. and Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. *Principles of Data Mining and Knowledge Discovery*, pages116-123.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137-142.

Lan, M., Tan, C. L., Su, J. and Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(4): 721-735.

Li, Y. H. and Jain, A. K. (1998). Classification of text documents. *The Computer Journal*, *41*(8): 537-546.

Quan, X., Wenyin, L. and Qiu, B. (2011). Term weighting schemes for question categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(5): 1009-1021.

Rak, R., Kurgan, L. and Reformat, M. (2005). Multi-label associative classification of medical documents from medline. In *Proceedings of Fourth International Conference on Machine Learning and Applications*, pages 177-186. IEEE.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1): 1-47.

Shum, S., Dehak, N., Dehak, R. and Glass, J. (2010). Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In *Proc. Odyssey*.

Soucy, P. and Mineau, G. W. (2005, July). Beyond TFIDF weighting for text categorization in the vector space model. In *International Joint Conference on Artificial Intelligence*, Vol. 19, pages 1130-1135. Lawrence Erlbaum Associates Ltd.

Tasci, S. and Gungor, T. (2008). An evaluation of existing and new feature selection metrics in text categorization. In *Proceedings of 23rd International Symposium on Computer and Information Sciences, ISCIS'08*, pages 1-6. IEEE.

Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H., and Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal*, *43*(3): 516-533.

Wen, C. Y. J. (2007). Text Categorization Based on a Similarity Approach. In *Proceedings of International Conference on Intelligent System and Knowledge Engineering*, Chengdu, China.

Xu, H., and Li, C. (2007). A Novel term weighting scheme for automated text Categorization. In *Proceedings of Seventh International Conference on Intelligent Systems Design and Applications, ISDA 2007*, pages 759-764. IEEE.