JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

Know Your Doctor: Topic Modeling and Sentiment Analysis Based Approach To Review Doctor

K. Kavya[1*], C. Sreejith[2]

[1,2]Calpine Labs, UVJ Technologies, Kochi, India

[*]*Corresponding Author: me.kavyakrishna@gmail.com*

Available online at: www.ijcseonline.org

Abstract— Nowadays people tend to search for doctors or firms through business review websites. They naturally opt for doctors that have the very best ratings and an outsized variety of reviews that support those high ratings. Hundreds or perhaps thousands of reviews will be given to the best-rated ones beneath their profiles, and comparing a high rated option to every alternative becomes a tedious task. This paper aims to address this issue by making a summarizer to analyze the doctors review by performing topic modeling using Latent Dirichlet Allocation(LDA) and Word2Vec based sentiment analysis. LDA is a standard Natural Language Processing (NLP) technique to determine topics from a large corpus. Word2vec based sentiment analysis is used to study people's opinions, attitudes and emotions towards a review. Word2vec is a neural network with two-layer that embeds the text corpus to a set of feature vectors of the words in the corpus. The reviews are taken from Yelp, an online rating website, of doctors across San Francisco. As a result of this study, a snapshot is created for each doctor with most dominant topics and the overall sentiment from their reviews.

Keywords— LDA, NLP, Sentiment Analysis, Topic Modeling, Word2Vec

I. INTRODUCTION

The Internet contains many web pages that look for reviews from online users about their experience of almost all services they get and all products they use. Although these reviews can give others a better elucidation about the service and its reliability, no user can spend their time reading the deluge of the reviews which are regularly fed into these online websites. To make it easy, several summarization systems were developed in the Natural Language Processing research arena.

Topic models are a collection of algorithms that provide an effective method to analyse massive volumes of unlabelled text in order to discover and determine underlying topic patterns in the form of a bag of words. Topic features and underlying word distribution of the topics can be found by optimizing the corpus and its pattern. The vast reviews can be shortened to a bag of topics which is distributed over words in vocabulary and can project the underlying category and type of the shorthand reviews. Thus users can easily go through the reviews which seems relevant to them instead of going through all possible options available online. This makes it easy for a user to understand the doctor with respect to their point of concern. For a person who thinks its better to have doctors near to their living place, won't prefer to go to affordable doctors far from their place. All these information won't be provided on any site. Reviews and suggestions from reviewers who visited the doctors are the only possibilities to learn about these underlying information. Analysing the vast reviews manually is not possible. Many review summarizer are developing currently in order to analyse reviews in different areas.

This paper aims at developing a model to integrate reviews on doctors, using topic modeling and sentiment analysis. In other words, for a given input review corpus, D, the system analyses massive volumes of unlabeled text in D in order to discover and determine underlying topic patterns in the form of a bag of words and assign sentiment scores to reviews under identified topics. It's a tedious task to go through thousands of reviews given to the different doctors and compare a high rated option to every alternative. This is applicable to cases where there is only one better-rated doctor. People also wish to know why others prefer this doctor and whether their concerns about doctors are addressed by the other reviewers. This again may take time.

Rest of the paper is organized as follows, Section I contains the introduction of Know Your Doctor, Section II contains the related work of topic modeling and sentiment analysis, Section III contains the system design and architecture of Know Your Doctor, Section IV contains the Results and Discussion and Section V concludes research work with future directions.

II. Related Work

Blei et al (2013) developed an unsupervised learning model Latent Dirichlet Allocation (LDA) under topic modelling class to generate word clusters under pre-specified topics from the massive collection of reviews [1].

Onan et al (2016) tested LDA in text sentiment classification by evaluating five classification algorithms (Naïve Bayes, support vector machines, logistic regression, radial basis function network and K-nearest neighbour algorithms) and five ensemble methods (Bagging, AdaBoost, Random Subspace, voting and stacking) on four sentiment datasets [9]. In order to analyze the working of LDA, the classification algorithms and ensemble algorithms were evaluated on datasets multi-domain sentiment, Irish-sentiment, review-polarity and reviews. They used classification accuracy and f-measure to evaluate the accuracy of each algorithm. As a result of that experiment, highest predictive performance for the datasets Irish-sentiment, multi-domain sentiment and review-polarity dataset was obtained through support vector machines compared to the others. For Reviews dataset, Stacking method yields the highest predictive performance. Ensemble learning methods yield generally better F-measure values compared to the weak learning algorithms [9].

Wang and Drach (2016) developed an unsupervised model that takes in the dataset of various artificial intelligence research paper and identifies the relevant topic in them [6]. It uses the dataset of research papers from JAIR.ORG, the Journal of Artificial Intelligence Research. The model is irrelevant without removing stopwords as it occurred most. They implemented this model using both K-means algorithm and LDA in order to compare the accuracy of both. Using K-means algorithm the topics generated were not of much relevance. The words in the articles seemed to tend towards only one or two topics. This was basically due to the outliers as the algorithm is susceptible to that [6]. On the other hand, LDA based model seemingly had better performance. But without removing stopwords LDA also performed poorly.

Wang et al (2016) introduced a new hybrid method to represent documents in a comprehensive way by incorporating Word2Vec with LDA to obtain relationships between documents and topics from LDA, as well as the contextual relationships from Word2Vec [10]. Support vector machines were used to set up the experiment with 20NewsGroup dataset [10].

Esposito et al (2016) conducted a comparative study on LDA and word2vec using the CompWHoB Corpus which is the abbreviation of Computational White House Press Briefings. In this paper, they start with Latent Dirichlet Allocation which is a standard topic modeling technique as the first step to generate hidden topics [11]. Further they use word2vec with properly preprocessed data on the basis of syntactic structure often termed as linguistic preprocessing. Results are poor when LDA is used alone and shows notable improvement when word embeddings generated by word2vec are used along efficient and task-focused linguistic pre-processing [11].

Sharma et al (2016) has done sentiment analysis of the patient reviews given to each doctor on different categories and used it to predict ratings for three main categories. They used a Convolutional Neural Network to perform sentiment analysis on reviews taken from "rated MD" website. They conclude that a Convolutional Neural Network(CNN) which has already pre-trained word embeddings trained using Adadelta based optimizer, categorical cross-entropy as a loss function and dropout technique as a regularizer gets better results for both binary and multiple classification problems [12].

Topic Modeling for Social Media Content paper by Rohani et al (2016) also experimented with creating an unsupervised topic modeling experiment in which LDA technique is used to get an idea about underlying probabilistic density of words and topics in social network corpus [13]. They used social media datasets in the domain of aviation and airport management.

A summarizer for unstructured online reviews was developed by Santosh and Vardhan (2015) using Resource Description Framework (RDF) to find the review features and sentiments of those reviews [4]. They included RDFS and OWL Ontology techniques along with machine learning algorithm [4]. Nuo Wang (2017) developed a similar summarizer for doctor reviews using LDA to detect the topics and used Vader sentiment analyser to retrieve the sentiments of doctor reviews [14]. Paul et al (2013) also developed a joint topic sentimental model to analyse online doctor reviews based on fractional LDA [22].

Word2vec and k-means clustering [3] were used over big data processing to classify data. On one hand, Word2Vec was used to find words in vocabulary whereas they clustered the similar words together and use the generated clusters to fit into a new data dimension so that the data dimension is decreased [3]. Word2vec and SVM perf [17] were together combined for sentiment analysis of Chinese comments on clothing products. Word2Vec also seemed a suitable technique for clustering but the combined model gave more accuracy about 90%.

An improved LDA model called GLDA [8] was developed for text classification to get more documents of relevant category by adding topic-category distribution parameter in the framework of LDA. Gibbs Sampling was employed in two datasets, Reuters-21578 dataset and Fudan Chinese Text Classification Corpus [8]. Sparse Constrained LDA (SC-LDA) [5] was developed by Yang et al (2015) to integrate prior knowledge to Latent Dirichlet Allocation and perform relatively better than other existing models for including word correlation and document label knowledge [5].

In this work, we are proposing LDA based Topic Modeling, combined with word2vec based Sentiment Analysis to generate effective doctor summary.

III. System design

This paper presents a model that is primarily designed to scrap review data from a website that contain several doctor reviews and ratings. The cleaned text data is topic modeled with LDA for generating topics. Among the generated topics, the most relevant 9 topics are manually selected and classified into general topics and doctor speciality topics. Their sentiment scores are obtained using Word2Vec and displayed along with topics. Review highlights of most positive and most negative comments are also shown. Figure. 1 shows the proposed system's flow chart.

A. Dataset

The basic step to start this project was to acquire the dataset for building a doctor review website. There was no independent doctor review dataset available. The only method to build the dataset is to scarp the reviews of doctors from existing online doctor review website. The first step is to look at the database of BetterDoctor [18] which gives ratings and details of doctors over a wide geographic area. It provides an API key through which data can be acquired from their website based on geographic location. Eventhough BetterDoctor website does not contain much reliable reviews, it gives the list of doctors having Yelp link [19], which is another website that provides reviews and ratings about a wide range of entities including doctors. The details and yelp link of 160 doctors around San Francisco were retrieved using API key from BetterDoctor. Finally, the Yelp pages of the above-listed doctors were downloaded for the dataset. Table 1 shows the sample reviews along with the corresponding doctor id.

B. Algorithms

1) LDA : It is a generative statistical model that extends Probabilistic Latent Semantic Analysis (PLSA), which is a technique from the family of topic models that model co-occurrence information under a probabilistic framework, in order to identify the hidden semantic structure of the data. Latent Dirichlet Allocation, a standard natural language processing (NLP) tool which determines underlying topics from an unstructured corpus automatically, is used to identify topics from the reviews to generate topics. As per LDA, each corpus, which consists of a collection of documents, is a blend of topics. Based on the probabilistic dirichlet distribution of the topics, LDA generates word vectors. Given some dataset containing sort of documents, LDA tries to work on finding which all topics will be blended to make up those documents within the starting place through backtracking. Figure. 2 shows a model of LDA. In mathematical terms, LDA can be seen as a matrix factorization technique. LDA represents document collection as a document-term matrix in the vector space which is further divided into document-topic matrix and topic-term matrix. LDA aims at improving the provided topic-word dirichlet distribution and document-topic dirichlet distribution given in the matrices.

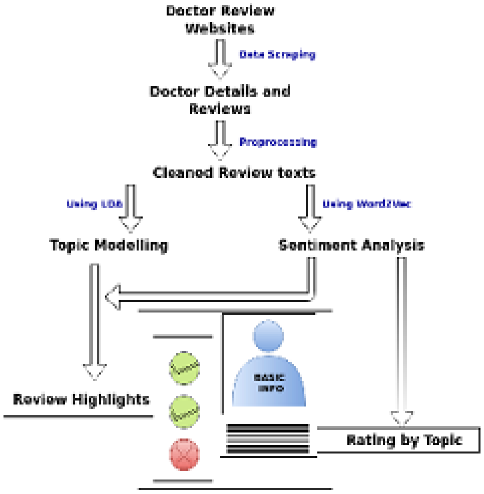
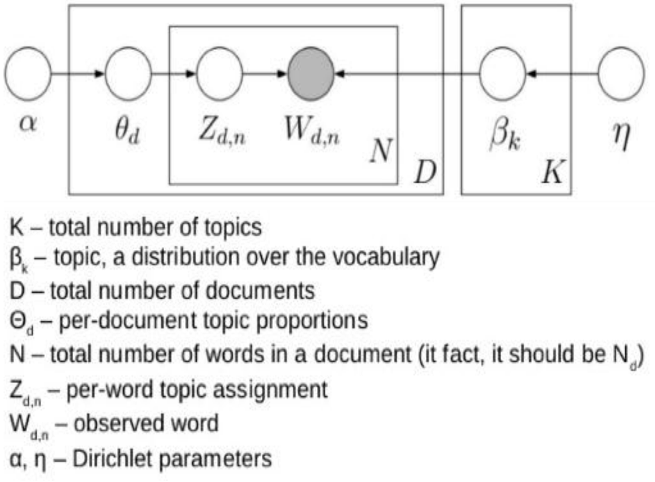


Figure 1. Know Your Doctor: Proposed Architecture



K – total number of topics

β_k – topic, a distribution over the vocabulary

D – total number of documents

Θ_d – per-document topic proportions

N – total number of words in a document (it fact, it should be N_d)

$Z_{d,n}$ – per-word topic assignment

$W_{d,n}$ – observed word

α, η – Dirichlet parameters

Figure 2. LDA Model [20]

For every document “d” in the corpus, LDA iterates through each word “w” throughout that document and fix the current topic-word assignment with a replacement assignment. For each word “w”, a new topic “z” is assigned with a probability P, where probability P = product(p1, p2), p1 and p2 are the two probabilities assigned to every topic in d.

Table 1. Reviews along with Doctor ID

Doctor ID	**Reviews**
1	"Dr. Shu possesses all the qualities of a great doctor- he is knowledgeable, empathetic, and personal. He listens to all my concerns without dismissing any of them as insignificant, and does not make me feel rushed during my appointment. "
1	"Dr. Shu has been my primary care doctor even though I live in San Jose. Having a doctor you can trust is hard to find, which is why i commute an extra 50 miles. Hoghly recommended."
2	"The WORST Doctor. Dr Ross is my primary but was not available after my auto accident. I had an AUTO ACCIDENT! She yelled at me for taking too much of her time! And she bad mouthed my primary Dr Ross"
3	"Dr Chiu Collins has returned my faith that there are plastic surgeons out there who truly care about their patients. She is polite, friendly, knowledgeable and is always prompt in her responses even outside of appointments. "
4	"Dr. Chen is very gentle and has my best interest in mind. Thank you"
7	"I was not impressed with Dr Lofquist. She wasn't rude but she didnt listen to one of my primary concerns & reason for seeing her, she dismissed it, and a simple blood test would have shown her that my concern was valid."

- p1 – the fraction of words occurring in the document "d" that is presently assigned to topic "z" (p(topic z / document d)).
- p2 – the fraction of mappings to the topic "z" from this word "w" in the entire documents (p(word w / topic z)).

A new topic which has the probability P is used for updating the present topic – word distribution. The existing word-topic distribution other than the present word is presumed to be correct by the model. Finally, the stable state is reached in which the document-topic dirichlet distribution and topic term dirichlet distribution are seemingly good after a number of iterations.

2) Topic Validation: Word2vec is a word embedding model introduced by Google which is meant to discover the semantic context of words with efficient linguistic preprocessing that also consist of some deep learning techniques [2]. Documents like reviews, press briefings, tweets, questions, answers or any other corpus are taken as input by the model and generates word embeddings in the vector space which may consist numerous features and dimensions by feature enrichment. Words that consist of similar meaning as well a similar context are embedded in the same vector space in word2vec model. The semantic structure of words is examined and looked into for further processing. Word2vec maps words in the vector space like a man is similar to a king in the context where a woman is similar to queen [2].

A set of features which are hard to categorise manually can be embedded automatically using word2vec. Word2vec is based on deep neural nets platform. This can be used to validate the generated topics. The same corpus is used for creating a topic distribution in vector space using word2vec. For the visualization of the given words in word2vec space t-distributed stochastic neighbor embedding (t-SNE) [7] dimensionality reduction method was used. The high-dimensional data was reduced to a two-dimensional space and plotted using scatter plot.

C. Sentiment Analysis

After identifying the topics, we assigned sentiment scores to sentences under identified topics. These topics generated are covered by reviews in our dataset. In order to score our criteria such as a doctor by the topics mentioned in their reviews, the sentiments of their reviews should be analysed. Word2vec along with SVC [21] is used for determining the sentiment of reviews. Labelled movie reviews were used as the training set for the classifier to classify the doctor reviews into positive and negative classes. The reviews were embedded into vector space using word2vec embedding algorithm. Thereafter average feature vector of each word embedding is found on the existing word2vec space. These average feature vectors are used by the classifier to classify the test data to sentiment classes.

IV. RESULTS AND DISCUSSION

This section describes the result of different algorithms and methods that were used. Also, discusses the output of all the modules of Know Your Doctor.

A. Topic Detection using LDA

Nine topics were detected using LDA algorithm and the topics distribution were given nine names manually: Payments, Positive comments, Appointment and office visits, Dental care, Women's health, Surgery, Allergy treatments, Skin procedures and Eyecare. Table 2 shows the assigned topic name along with the word distribution.

1) Topic Validation:

Word2Vec is used to validate the generated topics. The same corpus used for creating a topic distribution in vector space using word2vec. The most occurring words in each topic were plotted on word2vec space and plotted it using t-SNE dimensionality reduction method as shown in Figure. 3 and 4.

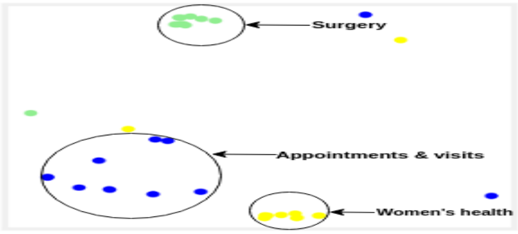


Figure 3. Topics 3, 5 and 6 Plotted

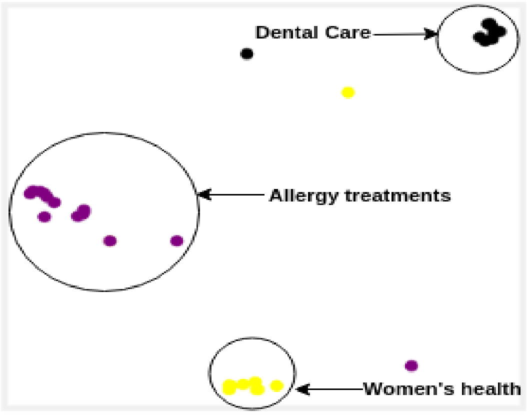


Figure 4. Topics 4, 5 and 7 Plotted

B. Sentiment Analysis

Using Word2Vec and SVC, sentiment of each review in the corpus is found. Word2Vec was tried along with many classifiers and SVC showed most accuracy among all others as shown in Table 3. Table 4 shows the sentiment score of each review calculated.

Table 3. Classifiers with their Training and Test Accuracy

Classifier	Training Accuracy	Test Accuracy
Random Forest	0.8055	0.7846
Gradient Boosting	0.9530	0.8365
Xgboost	0.8885	0.8337
Decision Tree	1.0	0.7022
SVM	0.8421	0.8317
SVC	0.8525	0.8420

Table 2. Generated Word Distribution And Assigned Topic Names

Topic Name	Word Distribution
1. Payments	“insur pay compani charg cover cost paid fee pocket provid amount money claim expens servic”
2. Positive comments	“great feel like staff realli friendli experi alway comfort nice recommend visit”
3. Appointment and office visits	“call appoint offic phone back get schedul week time messag receptionist result”
4. Dental care	“dentist teeth procedur root pain canal tooth clean fill wisdom”
5. Womens health	“babi pregnanc son deliv deliveri husband healthi child risk plan high daughter mother”
6. Surgery	“knee surgeon injuri shoulder mri month pain physic ankl heal bone recoveri therapi oper”
7. Allergy treatments	“allergi test shot asthma allerg allergist sinu disord reaction prick flew dust nasal cough infect dog inhal”
8. Skin procedures	“dermatologist mole face look scar acn cancer cosmet spot laser biopsi infect cream”
9. Eye care	“eye lasik glass contact prk laser face drop correct month facial wear look”

Table 4. Reviews Along With Doctor Id , Sentiment Class and Ratio

Reviews	ID	Class	Ratio
”Dr. Shu possesses all the qualities of a great doctor- he is knowledgeable, empathetic, and personal.”	1	1	0.53
”Dr Shu is awesome. I like his no nonsense advises on health, plus he spoke mandarin and that’s very helpful for my wife and my mother In-law”	1	1	0.56
”The WORST Doctor.I had an AUTO ACCIDENT! She yelled at me for taking too much of her time!”	2	0	0.59
”Thanks to Dr. Chiu my Botox injection for jaw slimming worked really well.”	3	1	0.59
”Dr. Chen is very gentle and has my best interest in mind. Thank you”	4	1	0.53

Reviews	ID	Class	Ratio
”Dr. Shu possesses all the qualities of a great doctor- he is knowledgeable, empathetic, and personal.”	1	1	0.53
”Dr Shu is awesome. I like his no nonsense advises on health, plus he spoke mandarin and that’s very helpful for my wife and my mother In-law”	1	1	0.56
”I was not impressed with Dr Lofquist. She wasn’t rude but she didnt listen to one of my primary concerns & reason for seeing her, she dismissed it.”	7	0	0.52

V. CONCLUSION and Future Scope

Know Your Doctor works as a review summarizer for understanding user's perception about doctors through their reviews and project a snapshot of each doctor from the given dataset which will give the user a quick glance at the doctor details. It gives the most discussed topics among the reviews of a particular doctor, their review summary, location and a word cloud using the prominent words in their reviews. From the existing models, LDA was found suitable for projecting topics whereas word2vec was used for finding the sentiment of the reviews.

The future scope includes using deep learning techniques to find the sentiment of reviews for improving accuracy. Extending the dataset by exploring more doctors will also make the summarizer platform more reliable and robust.

ACKNOWLEDGMENT

This work was fully funded by Calpine Labs, UVJ Technologies, Kochi, India. Also, this research was remarkably supported by Mr. Sreejith C, Data Scientist, Calpine Labs. We are also immensely grateful to Mr. Bijeesh Devassy, Project Manager, Calpine Labs and Dr. Asharaf.S, Associate Professor, IIITMK Trivandrum for sharing their pearls of wisdom with us during the course of this research. We also thank all of the colleagues from Calpine Labs, UVJ technologies, Kochi, India, who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] Blei, D. M., Ng, A. Y., \& Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning research, 3(Jan), 993-1022.

[2] Anon, (2018). [online] Available at: https://ahmedbesbes.com/sentiment-analysis-on-twitter-using-word2vec-and-keras.html [Accessed 13 Apr. 2018].

[3] Ma, L., \& Zhang, Y. (2015, October). Using Word2Vec to process big text data. In Big Data (Big Data), 2015 IEEE International Conference on (pp. 2895-2897). IEEE.

[4] Santosh, D. T., \& Vardhan, B. V. (2015). Obtaining feature-and sentiment-based linked instance RDF data from unstructured reviews using ontology-based machine learning. International Journal of Technology (2015) 2: 198, 2006.

[5] Yang, Y., Downey, D., \& Boyd-Graber, J. (2015). Efficient methods for incorporating knowledge into topic models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 308-317).

[6] Wang, Stephanie, \& Drach, Max. (2017). Latent Dirichlet Allocation for Identifying Topics in AI.

[7] Maaten, L. V. D., \& Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.

[8] Zhao, D., He, J., \& Liu, J. (2014, April). An improved LDA algorithm for text classification. In Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on (Vol. 1, pp. 217-221). IEEE.

[9] Onan, A., Korukoglu, S., \& Bulut, H. (2016). LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. Int. J. Comput. Linguistics Appl., 7(1), 101-119.

[10] Wang, Z., Ma, L., \& Zhang, Y. (2016, June). A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec. In Data Science in Cyberspace (DSC), IEEE International Conference on (pp. 98-103). IEEE.

[11] Esposito, F., Corazza, A., \& Cutugno, F. (2016). Topic Modelling with Word Embeddings. CLiC it, 129.

[12] Sharma, R. D., Tripathi, S., Sahu, S. K., Mittal, S., \& Anand, A. (2016). Predicting online doctor ratings from user reviews using convolutional neural networks. International Journal of Machine Learning and Computing, 6(2), 149.

[13] Rohani, V. A., Shayaa, S., \& Babanejaddehaki, G. (2016, August). Topic modeling for social media content: A practical approach. In Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on (pp. 397-402). IEEE.

[14] " GitHub. (2018). nuwapi/DoctorSnapshot. [online] Available at: https://github.com/nuwapi/DoctorSnapshot [Accessed 12 Apr. 2018]."

[15] "Pypi.python.org. (2018). gmplot 1.0.5 : Python Package Index. [online] Available at: https://pypi.python.org/pypi/gmplot/1.0.5 [Accessed 12 Apr. 2018]."

[16] "Pypi.python.org. (2018). wordcloud 1.4.1 : Python Package Index. [online] Available at: https://pypi.python.org/pypi/wordcloud [Accessed 12 Apr. 2018]."

[17] "Zhang, D., Xu, H., Su, Z., \& Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 42(4), 1857-1863."

[18] "Betterdoctor.com. (2018). BetterDoctor :: BetterDoctor - The Origin of Accurate Provider Data. [online] Available at: https://betterdoctor.com [Accessed 9 Apr. 2018]."

[19] "Yelp. (2018). Yelp. [online] Available at: https://www.yelp.com/ [Accessed 9 Apr. 2018]."

[20] "Alexis Perrier - Data Science. (2018). Segmentation of Twitter Timelines via Topic Modeling. [online] Available at: https://alexisperrier.com/nlp/2015/09/16/segmentation_twitter_timelines_lda_vs_ \newline lsa.html [Accessed 9 Apr. 2018]."

[21] http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForest\newline Classifier.html

[22] Paul, M. J., Wallace, B. C., & Dredze, M. (2013, June). What affects patient (dis) satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model. In AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI.