How do patients talk about Tinnitus

Paras Multani
paras.multani@st.ovgu.de

Rajeev Motwani
rajeev.motwani@st.ovgu.de

Kantharaju CN
kantharaju.c@st.ovgu.de

Prakruthi Shivanna
prakruthi.shivanna@st.ovgu.de

March 28,2016

Abstract Tinnitus is becoming increasingly prevalent because of increased daily noise levels, including those caused by unrestrained use of headphones and stereos etc. The majority of people who suffer from tinnitus find it very disturbing and uncomfortable. So far, there is no scientific proven treatment or cure for tinnitus but there are ways to ease it and reduce its impact. Patients post their problems about tinnitus on "tinnitustalk.com". These posts include their problems, symptoms and their knowledge about the treatment of tinnitus. Often, providing a treatment for a particular symptom seems difficult. The objective in our project is to ascertain how do people talk about tinnitus. We tried to group and classify the posts, also we attempted to identify topics in these posts in order to help medical experts in making better decisions regarding the treatment of tinnitus. The paper describes the various tasks involved in our analysis, including the posts collection activity and various mining algorithms applied on these posts.

1 Introduction

Tinnitus is the perception of noise or ringing in the ears. Its a symptom of an underlying condition, such as age-related hearing loss, ear injury or a circulatory system disorder. It involves the annoying sensation of hearing sound when no external sound is present [1]. Tinnitus can significantly affect the quality of life, people suffering from tinnitus may experience stress, sleep

problems, depression, anxiety and fatigue. For many people, tinnitus can improve with treatment and in some cases treatments can reduce or mask the noise, making tinnitus less noticeable.

In our study, we have performed various data mining activities which involve clustering, classification and topic modelling in order to understand how do patients talk about their illness i.e tinnitus in this case. The text data has been collected in the form of posts from the 'Introduce Yourself' and 'Treatments' section of the online 'Tinnitus Talk Support Forum'. In the introduction section, new users usually express their history with tinnitus and share their problems in order to get some help and advice from people who may have faced similar symptoms in the past. In the treatments section, people usually talk about the drugs and supplements which they have consumed and various therapies that could prove effective in certain cases. The text mining in our analysis is based on the posts from last five years, close to 2500 posts in the introduce section and around 72 posts in the treatments section. The remainder of the paper is structured as follows: Section 2 provides the details on how the data was gathered for the task. Section 3 provides the concise implementation idea about the used methods for clustering, classification and topic modelling on the crawled dataset. Section 4 describes the evaluation of the used methods and presents statistics. Section 5 remarks on the project plan, the tools used, problems faced and thus concludes the paper.

2 Data Collection

Crawling the Web It refers to downloading pages from the web. Crawlers are widely used nowadays, for extracting large datasets from the web, but the design of a good crawler poses many challenges since the web is being constantly updated. We have implemented a web crawler using 'jsoup' library in Java. It starts with the 'Introduce Yourself' and 'Treatments' as the seed URLs to be visited first. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to be visited. Then, there is a recursive loop to visit all the pages till date for both sections. From the source pages, the values for the tags - user, title, publish date and description are extracted and the results are stored in MySQL database schema. A sample of data stored in the database is displayed in figure 1 below.

userid	title	publishd	description
.bill	Hello	Feb 1...	Had a case of flu Then ears felt plugged up Then the ringing started sto...
1MW	SSNHL Tinnitus Hyperacusis	Nov ...	Hello Some day after afternoon sleep when i wake up noticed my ear rin...
1Regcabguy$$	Newbie Here	Jan 3...	Hello New tinnitus sufferer here Injured my hearing recently with firear...
2131e	New to This How Loud Is Your Tinni...	Sep 1...	Hi everyone Im so grateful to have found all of you guys here on this fo...
2Fight	Aneurysm Survivors Story — Onset...	Jun 2...	I just had my 2nd Anniversary of Aneurysm rupture in the right PICA of ...
2lo	37 Years Old Just Got Diagnosed wi...	Aug ...	Hi all New to this situation Not a happy camper I guess you could say 37...
2LoudnHear	Its 2LoudnHear	Aug ...	Im treading water in a sea of sound
2Nine	A New Tinnitus Kid in Town	Sep 1...	Hello there my name is Pavel and I'm a musician I play the guitar not in a...
400runner	AntiInflammatory Induced	Nov ...	Hi Everyone I had a back injury at work for which I was prescribed the a...

Figure 1 : Sample of database schema

3 Data Mining Activities

Posts Clustering Its an idea similar to the concept of document clustering which refers to extracting interesting and non-trivial patterns or knowledge from unstructured text data [2]. In our task, clustering involves natural language processing, text analysis and categorization. We deal with corpus based computational linguistics over the large collection of posts in order to discover useful patterns. At first, data cleaning is performed using NLP tasks like stopword removal, eliminating unnecessary punctuation and whitespaces on the corpus. This is done using the text mining package 'tm' in RStudio. Then, a term-document matrix is developed for the cleaned dataset in which the tf-idf vector weight scheme is used to determine the values in the matrix. Finally, euclidean distance is normalized and k-means clustering is performed on the matrix. Figure 2 below, displays the percentage of posts in each cluster for both the sections 'Introduce Yourself' and 'Treatments'.

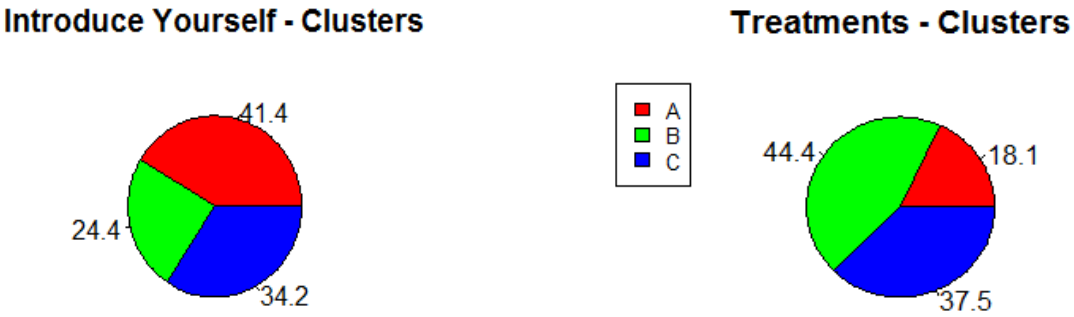


Figure 2 : Proportional representation of posts in k = 3 clusters

Text Classification While the clustering is being performed for the posts, a text classifier is developed simultaneously to assign a post to one of the following classes i.e positive, negative or neutral. Text classification has an important role to play, especially with the increase in day-to-day incoming posts by patients. The notion of classification in our approach is very general and we have implemented a simple classifier in Java using SentiWordNet. It is a lexical resource for text mining and sentiment analysis based on WordNet, an english lexical database, SentiWordNet associates every WordNet synonym collection to three numerical scores positive, negative or neutral. Each of the three scores is normalized between 0 and 1, and their sum is equal to 1. A final score is calculated for every post by summing up the score of words present in the post and based on this, each post is assigned either to positive, negative or neutral class. If the score ranges between -1 to 1, it is termed as neutral post, if the score goes more than 1, the post belongs to positive class or else if the score goes below -1, the post becomes negative [3]. Based on the scores in each of the three polarities, the posts are ranked in the decreasing order of their scores. Finally, we identify the dominant polarity in each cluster for both 'Introduce Yourself' and 'Treatments' section. Figures 3 and 4 below, show the count of polarities in each cluster for 'Introduce Yourself' and 'Treatments' sections respectively.

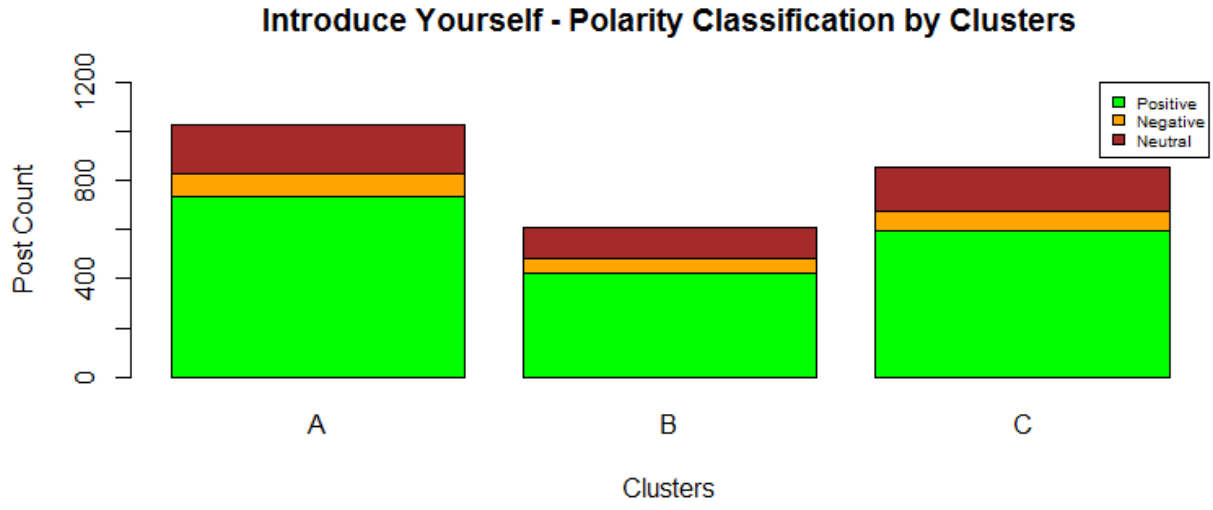


Figure 3 : Classification in Introduce Yourself dataset

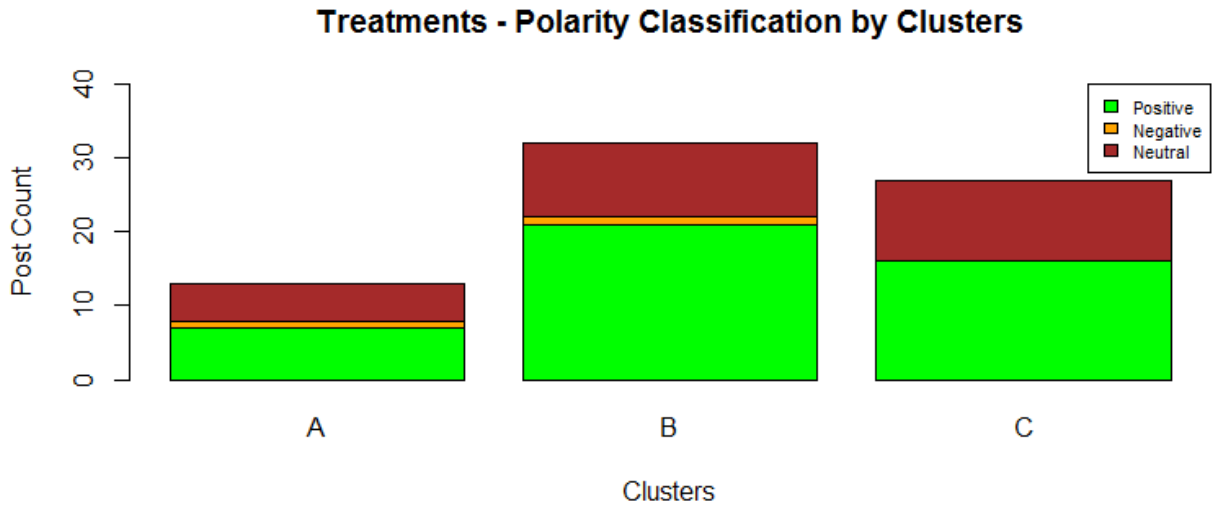


Figure 4 : Classification in Treatments dataset

Topic Modelling Since large amounts of posts are coming in daily on the online support forum for tinnitus, it becomes difficult to access what one is looking for and hence, we perform topic modelling which helps us to organize, understand and summarize large collections of posts. Topic Models are based on the idea that posts are a mixture of topics, where a topic is a probability distribution over words [4]. It is a text mining tool for discovery of abstract topics that occur in a collection of posts. In our approach, first a term-document matrix is created for every cluster and then using the log-likelihood method in R, we determine the best number of topics for the matrix i.e 'k'. Then, we use the Latent Dirichlet Allocation (LDA) technique to obtain the topics that are learnt using Gibbs sampling. In the LDA model, each word in the post is attributable to one of the post's topics [5]. Finally, we output the terms with highest probabilities for every topic within the cluster. Figures 5 and 6 below, describe these terms for three topics within each cluster in 'Introduce Yourself' and 'Treatments' section respectively. From the data in these figures, we try to abstractly describe each topic within a cluster and finally based on the abstract definitions we give every cluster a name.

Introduce_Clusters	Topic 1	Topic 2	Topic 3
A	vertigo	loud	plugs
	hissing	neck	vertigo
	annoying	hyperacusis	neck
B	perforation	whistling	sudafed
	beep	pregnancy	lexapro
	mucus	fever	earache
C	humming	dose	siren
	pulsing	whistling	hiss
	whistling	intensity	block

Figure 5 : Topics in Introduce Yourself Clusters

Cluster A : Abstract Topic 1 Name {Disease,Sound}
Abstract Topic 2 Name {Disease,Sound}
Abstract Topic 3 Name {Disease}
Cluster B : Abstract Topic 1 Name {Infection}
Abstract Topic 2 Name {Infection,Sound}
Abstract Topic 3 Name {Decongestants,Infection}
Cluster C : Abstract Topic 1 Name {Sound}
Abstract Topic 2 Name {Quantity,Sound}
Abstract Topic 3 Name {Sound}

Tretment_Clusters	Topic 1	Topic 2	Topic 3
A	cbt	drug	retigabine
	biloba	antidepressants	drug
	bppv	elacin	neurofeedback
B	gaba	otic	estimme
	vascular	ganglion	neurostimulator
	neuromonics	otoharmonics	zinc
C	stem	epilepsy	carbamazepine
	ringing	loudness	ringing
	melatonin	magnesium	noises

Figure 6 : Topics in Treatments Clusters

Cluster A : Abstract Topic 1 Name {Drugs,Therapy}
Abstract Topic 2 Name {Drugs}
Abstract Topic 3 Name {Anticonvulsant,Drugs}
Cluster B : Abstract Topic 1 Name {Devices,Supplements}
Abstract Topic 2 Name {Disorder,Supplements}
Abstract Topic 3 Name {Treatment,Supplements}
Cluster C : Abstract Topic 1 Name {Sound,Treatment}
Abstract Topic 2 Name {Health Condition,Sound}
Abstract Topic 3 Name {Sound,Drugs}

From the above data, we can infer the cluster names as below, which suggest the most popular subject being talked about in that cluster.
Clusters in Introduce Yourself
Cluster A - Disease
Cluster B - Infection
Cluster C - Sound

Clusters in Treatments
Cluster A - Drugs
Cluster B - Supplements
Cluster C - Sound

4 Evaluation and Statistics

To assess the quality of clustering algorithm with respect to different values of k i.e. the number of clusters, we have made use of an optimal parameter called Silhouette co-efficient. The silhouette value is a measure of how similar a post is to its own cluster compared to other clusters. The value for this parameter ranges from -1 to 1 where higher value indicates that the post is well matched to its own cluster and poorly matched to the remaining clusters. We have calculated the silhouette using euclidean distance metric and 'factoextra' library in R.

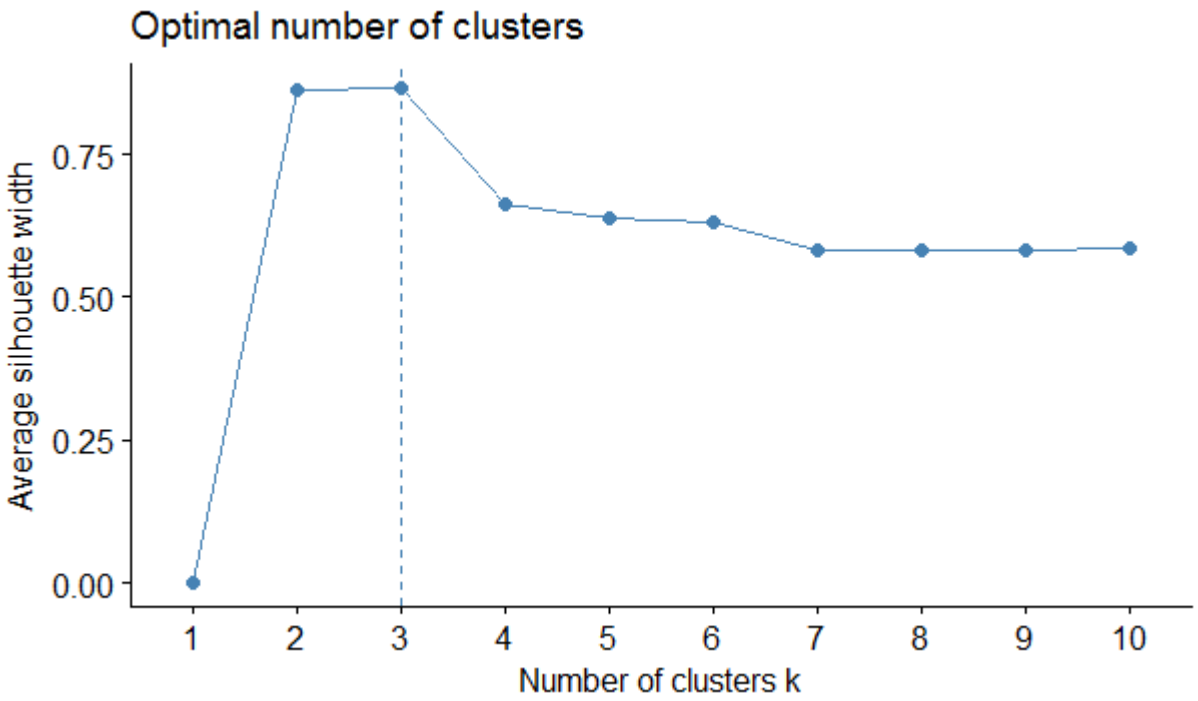


Figure 7 : Silhouette Plot for Introduce Yourself dataset

The figures 7 and 8 present the evaluation graph in terms of number of optimal clusters to be generated for the posts in 'Introduce Yourself' and 'Treatments' respectively. In both cases, we get the best k value as 3.

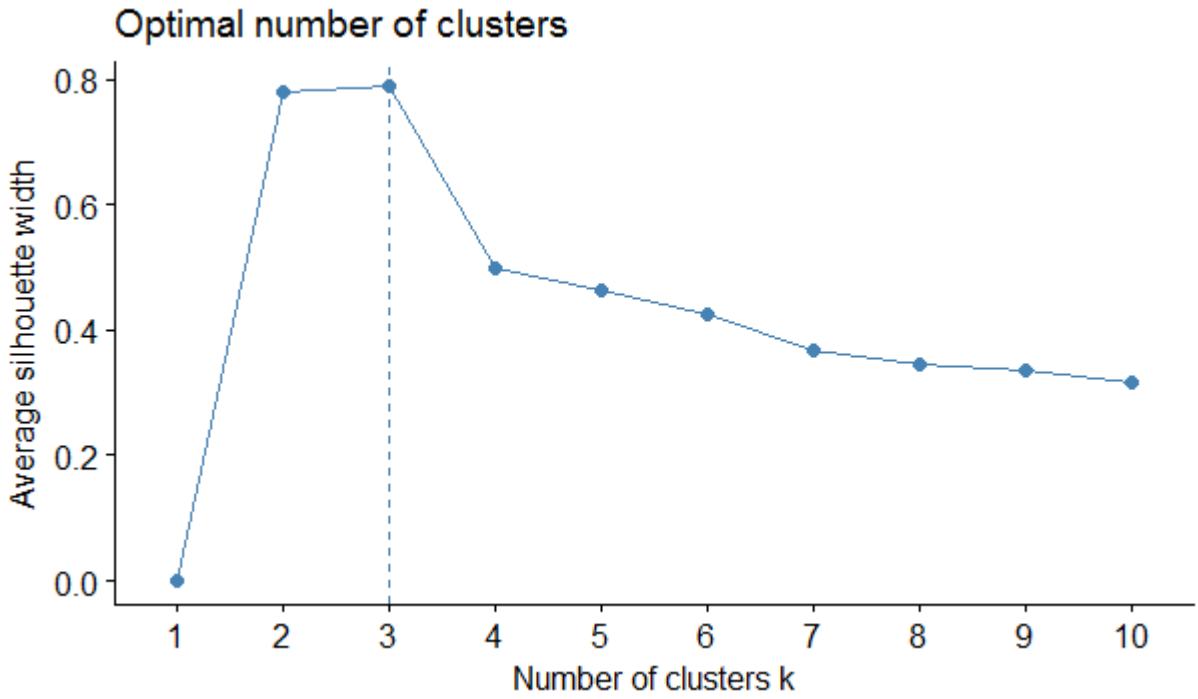


Figure 8 : Silhouette Plot for Treatments dataset

Apart from this, we were given a list of entity descriptors during the start of our project, for which basic statistical computations had to be done. Below, we present the graphs 9 and 10 for the same. Also, we monitored the variation of posts for the year 2016, in both 'Introduce Yourself' and 'Treatments' section in figures 11 and 12, and the surprising note here is that only four discussions have been posted throughout the year for 'Treatments' which highlights the fact that either people are less aware about the medication for tinnitus or they are not sure about which drug to be consumed for a particular symptom.

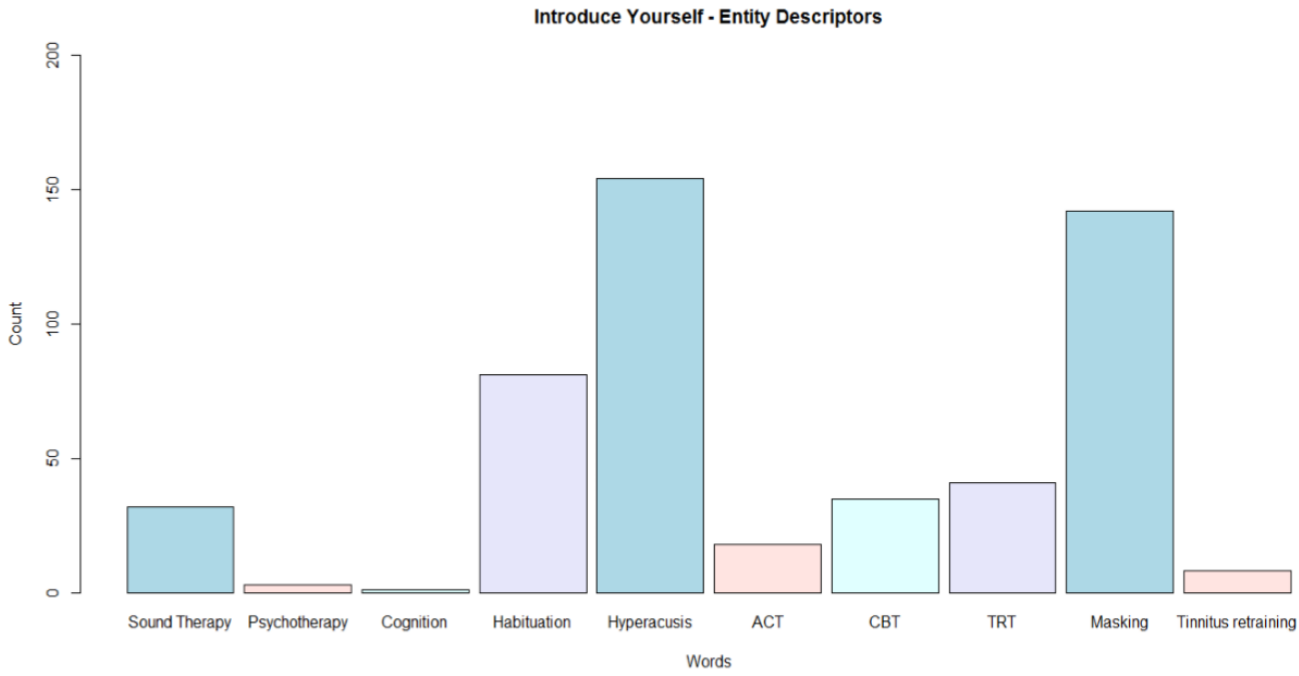


Figure 9 : Entity descriptor statistics for Introduce Yourself dataset

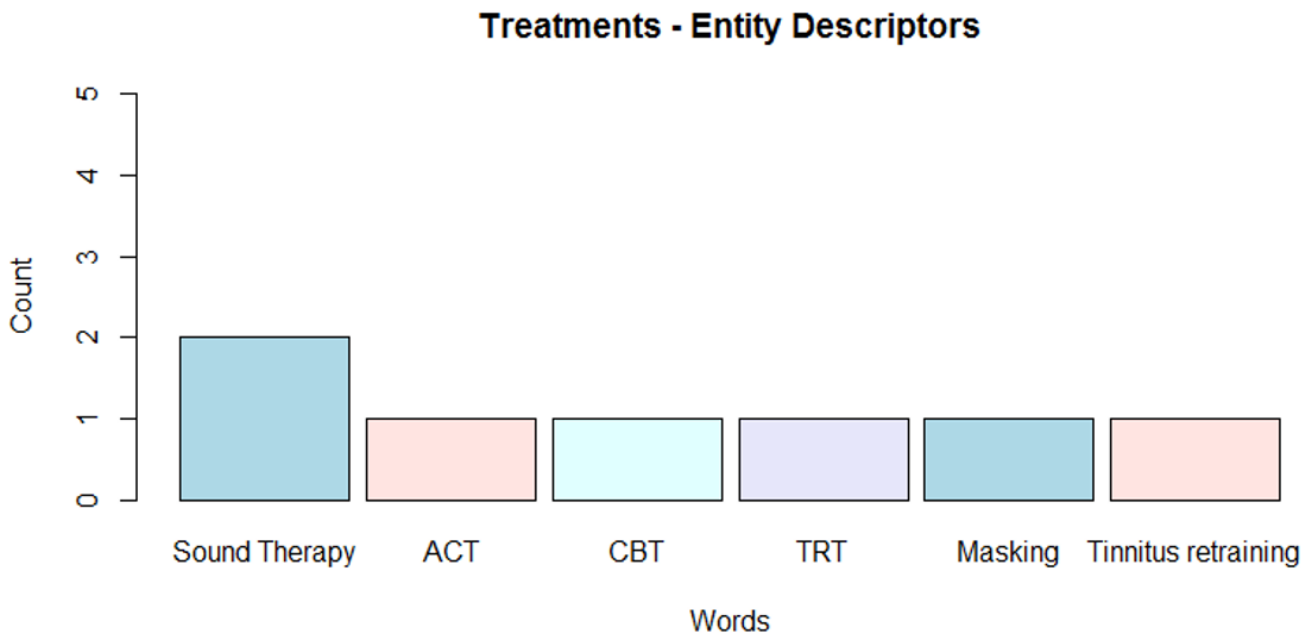


Figure 10 : Entity descriptor statistics for Treatments dataset

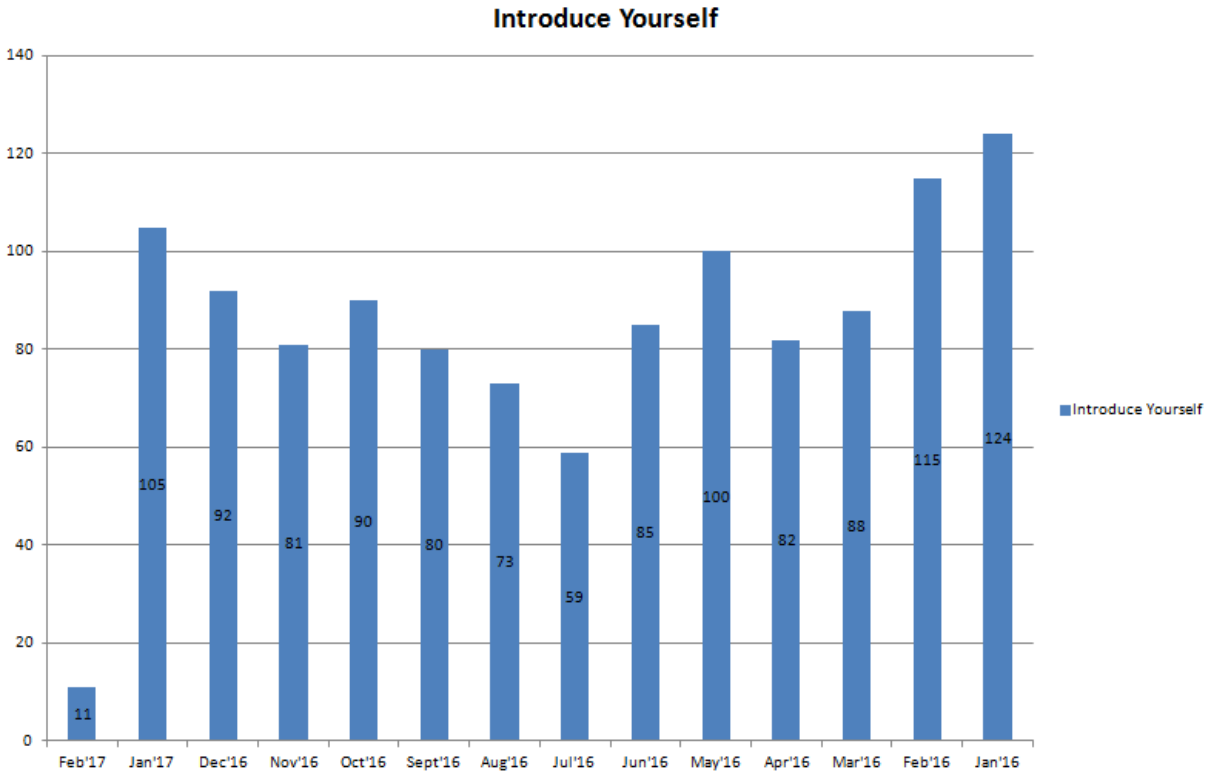


Figure 11 : Introduce Yourself postings count for the year 2016

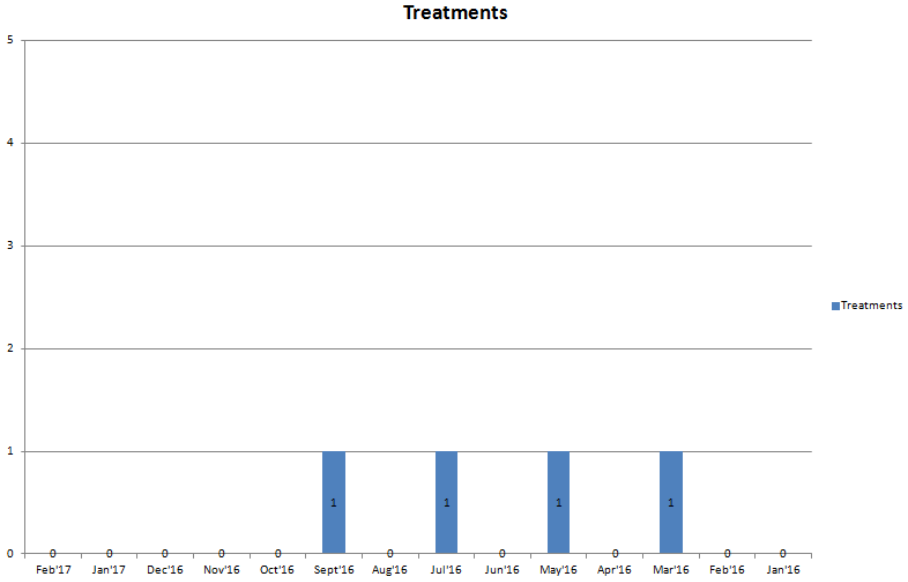
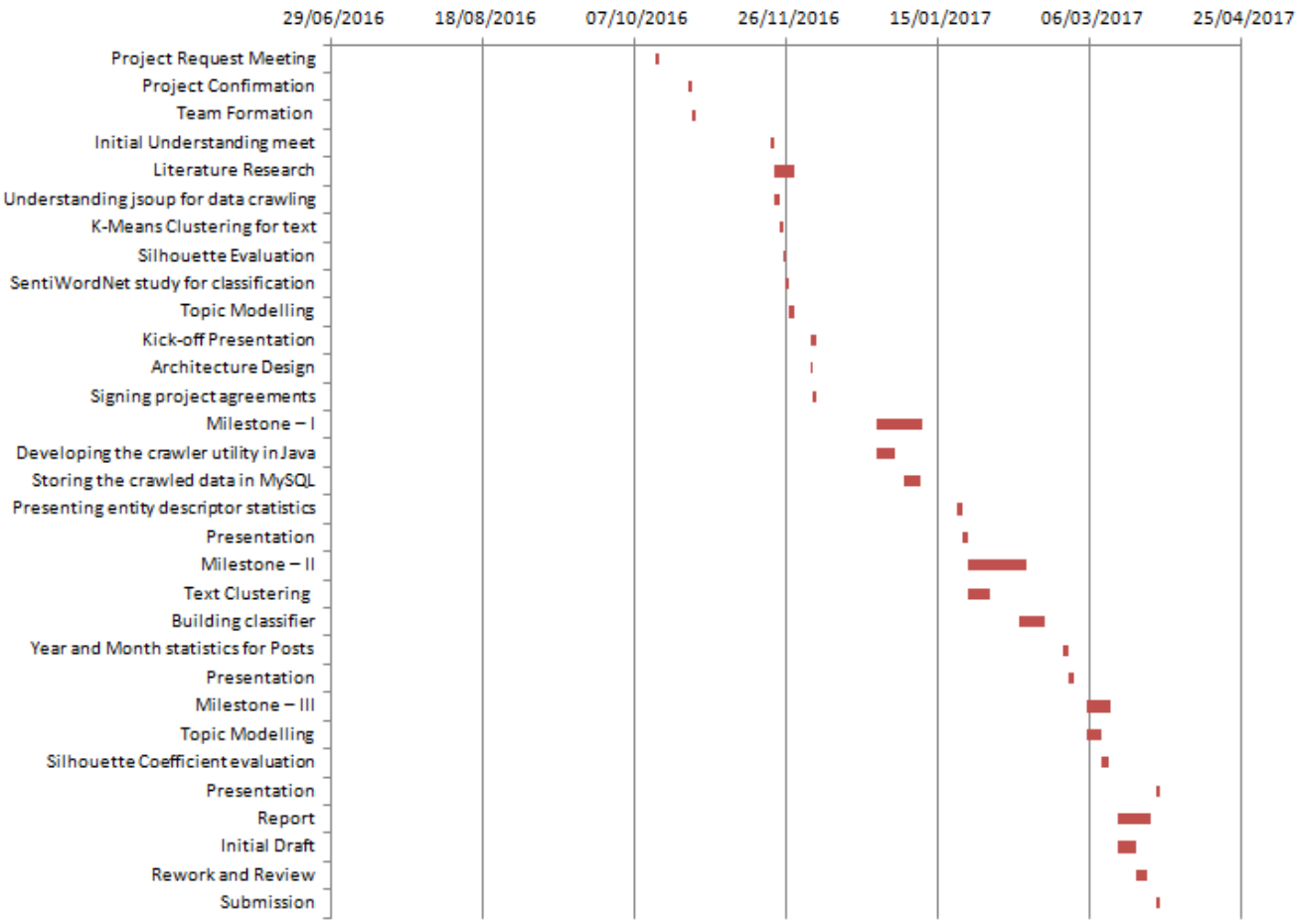


Figure 12 : Treatments postings count for the year 2016

5 Project Plan and Technical Specifications

Task	Start Date	Duration	End Date	Person Responsible
Project Request Meeting	14-10-2016	1	14-10-2016	Paras,Kantharaju
Project Confirmation	25-10-2016	1	25-10-2016	Prof. Myra Spiliopoulou
Team Formation	26-10-2016	1	26-10-2016	Paras
Initial Understanding meet	21-11-2016	1	21-11-2016	Prof. Myra Spiliopoulou
Literature Research	**22-11-2016**	**7**	**28-11-2016**	**All team members**
Understanding jsoup for data crawling	22-11-2016	2	23-11-2016	
K-Means Clustering for text	24-11-2016	1	24-11-2016	
Silhouette Evaluation	25-11-2016	1	25-11-2016	
SentiWordNet study for classification	26-11-2016	1	26-11-2016	
Topic Modelling	27-11-2016	2	28-11-2016	
Kick-off Presentation	**4-12-2016**	**2**	**5-12-2016**	
Architecture Design	4-12-2016	1	4-12-2016	Kantharaju
Signing project agreements	5-12-2016	1	5-12-2016	All team members
Milestone – I	**26-12-2016**	**15**	**24-1-2017**	
Developing the crawler utility in Java	26-12-2016	6	31-12-2016	Paras,Kantharaju
Storing the crawled data in MySQL	4-1-2017	5	8-1-2017	Rajeev,Prakruthi
Presenting entity descriptor statistics	21-1-2017	2	22-1-2017	Paras,Rajeev
Presentation	23-1-2017	2	24-1-2017	Kantharaju,Prakruthi
Milestone – II	**25-1-2017**	**19**	**28-2-2017**	
Text Clustering	25-1-2017	7	31-1-2017	Paras,Rajeev
Building classifier	11-2-2017	8	18-2-2017	Kantharaju
Year and Month statistics for Posts	25-2-2017	2	26-2-2017	Prakruthi
Presentation	27-2-2017	2	28-2-2017	Prakruthi
Milestone – III	**5-3-2017**	**8**	**28-3-2017**	
Topic Modelling	5-3-2017	5	9-3-2017	Paras,Rajeev
Silhouette Coefficient evaluation	10-3-2017	2	11-3-2017	Kantharaju,Prakruthi
Presentation	28-3-2017	1	28-3-2017	Paras,Rajeev
Report	**15-3-2017**	**11**	**28-3-2017**	
Initial Draft	15-3-2017	6	20-3-2017	Paras
Rework and Review	21-3-2017	4	24-3-2017	Rajeev,Prakruthi,Kantharaju
Submission	28-3-2017	1	28-3-2017	All team members



Tools and Technologies
The tools used in the development of the complete task include
Java programming
Python programming
R-Studio
MySQL
MS-Excel

Problems The only issue faced while working on the project was generating large matrix from 'Introduce Yourself' dataset, as the team members had maximum 4gb RAM configurations which was not sufficient for such huge data. The problem was resolved by performing data reduction and as-

signing maximum system memory to R.

References

[1]. Papadakis MA, et al., eds. Ear, nose, and throat disorders. In: Current Medical Diagnosis and Treatment 2015. 54th ed. New York, N.Y.: The McGraw-Hill Companies; 2015.

[2]. Tan, Ah-Hwee. "Text mining: The state of the art and the challenges." Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. Vol. 8. 1999.

[3]. Agrawal, Shaishav. "Using syntactic and contextual information for sentiment polarity analysis." Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human. ACM, 2009.

[4]. Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." Handbook of latent semantic analysis 427.7 (2007): 424-440.

[5]. D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993-1022, 2003