A MODEL FOR IMPLEMENTING HEALTH CARE SERVICE TICKET CLASSIFICATION USING NLP

D. Anil Kumar[1], Sudheer Babu Punuri[2], P.S.Prema Kumar [3], A V S Pavan Kumar [4], Dr Mula Malyadri[5]

[1]Associate Professor, Department of CSE, GIET University, Gunupur
[2]Assistant Professor, Dept of CSE, GIET University,Gunupur
[3]Associate Professor, Dept of Mechanical,Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh
[4]Assistant professor, Department of CSE,GIET University, Gunupur
[5]Associate Professor, Department of CSE,CMR Technical Campus, Hyderabad, Telangana, India

ABSTRACT:

Nowadays getting doctor appointment is very difficult task The Heath Care Services range from basic medical diagnostics to critical emergency services offering a ticketing system for all the telephonic calls received across all the departments.Calls to the provider can be for New Appointment, Cancellation, Lab Queries, Medical Refills, Insurance Related, and General Doctor Advise etc.The Tickets have the details of Summary of the call and description of the calls written by various staff members with no standard text guidelines. We investigate to see if, based on the Text in the 'Summary' and 'Description' of the call, the ticket can be automatically classified to Appropriate Category (out of 6 Categories) and Subcategories (Out of 21 Sub Categories) with good accuracy using Machine learning approach.

We use the bag of words approach to solve the problem. Further we would try different data feature representation methods using Document Term Matrix (tf, tf-idf, binary-tf,) and different Machine learning algorithms and see which performs better, and try to reason why one works better than the other.

1. INTRODUCTION:

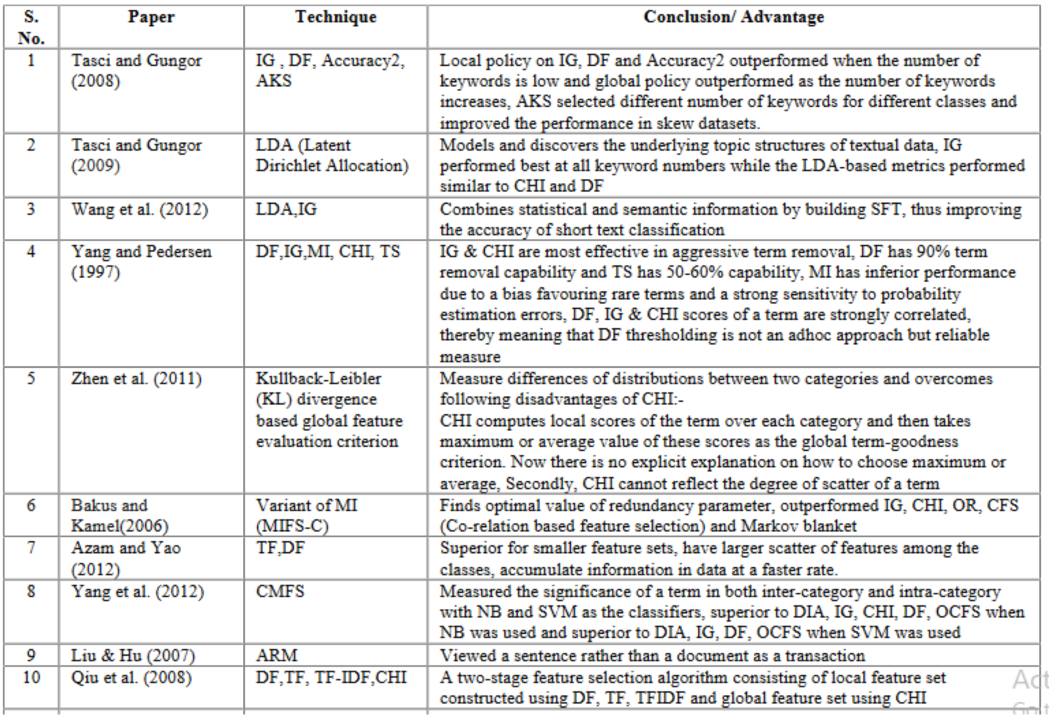
Text Classification is an old problem. The methods have changed over time, from linguistic scholars pouring over documents - to rule based approaches which make use of human intuition and are developed by experienced language engineers -to a purely statistical computer-based approach. The large amount of data in the form of text available on internet and the need of organizing it has generated and progressively intensified the interest in automatic text categorization. A widely-used research approach to this problem is based on machine learning techniques: an inductive process which builds a classifier by automatically learning the characteristics of the categories. Machine learning is more portable and less labour-intensive than manual definition of classifiers.

Text Classification has been applied in many applications such as authorship detection, plagiarism detection, fraud detection in criminology and Security domain, Spam Detection, automated bucketing of service tickets etc .These days text categorization is a discipline at the crossroads of ML and IR, and it claims a number of characteristics with other tasks like information/ knowledge pulling from texts and text mining. Automatic Text Categorization can be application to improve the efficiency of Text Retrieval on the internet.

The most prominent features of the machine learning approach to text classification, include data preparation, attribute extraction and selection, learning algorithms and kernel methods and fine tuning them, performance measures, and availability of training corpora. Bag of Words approach is used and the words become the features for applying machine learning algorithms. The n-gram approach provides a natural representation to keep into account the order of words(context) in the input text. Unlike bag of words plus stemming, the n-gram approach can be applied without previous knowledge of the language of the input. Most modern approaches on text classification use diverse methods such as decision trees, support vector machines, neural networks, KNN and Bayesian classifiers depending on the size and type of the problem. Decision Tree methods suffer if the number of features is very high. Feature selection becomes important for applying decision trees. SVM methods have been shown to be excellent on text classification tasks both theoretically and experimentally with less feature selection. Recently, Multilayer Perceptron, Recurrent Neural Networks and Convolution Neural Networks have been used for text classification problem. They have given good accuracy without much pre-processing compared to linear Machine learning methods.

In this particular project we are using linear models such as SVM, KNN, Decision Trees and Random Forests to build the model.

2. LITERATURESURVEY:



S. No.	Paper	Technique	Conclusion/ Advantage
1	Tasci and Gungor (2008)	IG , DF, Accuracy2, AKS	Local policy on IG, DF and Accuracy2 outperformed when the number of keywords is low and global policy outperformed as the number of keywords increases, AKS selected different number of keywords for different classes and improved the performance in skew datasets.
2	Tasci and Gungor (2009)	LDA (Latent Dirichlet Allocation)	Models and discovers the underlying topic structures of textual data, IG performed best at all keyword numbers while the LDA-based metrics performed similar to CHI and DF
3	Wang et al. (2012)	LDA,IG	Combines statistical and semantic information by building SFT, thus improving the accuracy of short text classification
4	Yang and Pedersen (1997)	DF,IG,MI, CHI, TS	IG & CHI are most effective in aggressive term removal, DF has 90% term removal capability and TS has 50-60% capability, MI has inferior performance due to a bias favouring rare terms and a strong sensitivity to probability estimation errors, DF, IG & CHI scores of a term are strongly correlated, thereby meaning that DF thresholding is not an adhoc approach but reliable measure
5	Zhen et al. (2011)	Kullback-Leibler (KL) divergence based global feature evaluation criterion	Measure differences of distributions between two categories and overcomes following disadvantages of CHI:- CHI computes local scores of the term over each category and then takes maximum or average value of these scores as the global term-goodness criterion. Now there is no explicit explanation on how to choose maximum or average, Secondly, CHI cannot reflect the degree of scatter of a term
6	Bakus and Kamel(2006)	Variant of MI (MIFS-C)	Finds optimal value of redundancy parameter, outperformed IG, CHI, OR, CFS (Co-relation based feature selection) and Markov blanket
7	Azam and Yao (2012)	TF,DF	Superior for smaller feature sets, have larger scatter of features among the classes, accumulate information in data at a faster rate.
8	Yang et al. (2012)	CMFS	Measured the significance of a term in both inter-category and intra-category with NB and SVM as the classifiers, superior to DIA, IG, CHI, DF, OCFS when NB was used and superior to DIA, IG, DF, OCFS when SVM was used
9	Liu & Hu (2007)	ARM	Viewed a sentence rather than a document as a transaction
10	Qiu et al. (2008)	DF,TF, TF-IDF,CHI	A two-stage feature selection algorithm consisting of local feature set constructed using DF, TF, TFIDF and global feature set using CHI

Over the decades text classification problem studied widely for various real world applications [1-8]

The widely used techniques for extracting the feature are (TF-IDF), Term Frequency (TF) [9], Word2Vec [10], and Global Vectors for WordRepresentation (GloVe) [11].

In this paper ensemble learning based techniques are used for text analysis which is providing good accuracy[12].

The simplest classification techniques is linear regression which is widely used to solve the data mining problem but it will work effectively when the input data is text[18].

Naïve Bayes Classifier is used more frequently because it is computationally low cost and also memory requirements are very less [19].

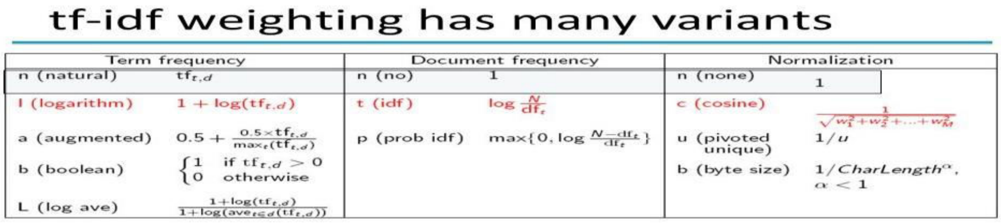
Tree-based classifiers such as decision tree and random forest have also been studied withrespect to document categorization [23].

3. METHODOLOGY:

3.1 Bag of Words Representation of Text:

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model is commonly used in methods of document classification where the frequency of occurrence of each word is used as a feature for training a classifier.

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes for determining the value that each entry in the matrix should take. Generally tf-idf is the most used scheme. There are various other schemes. with to see which give the best results.



tf-idf weighting has many variants

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log\frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0,\log\frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

4. ARCHITECTURE:

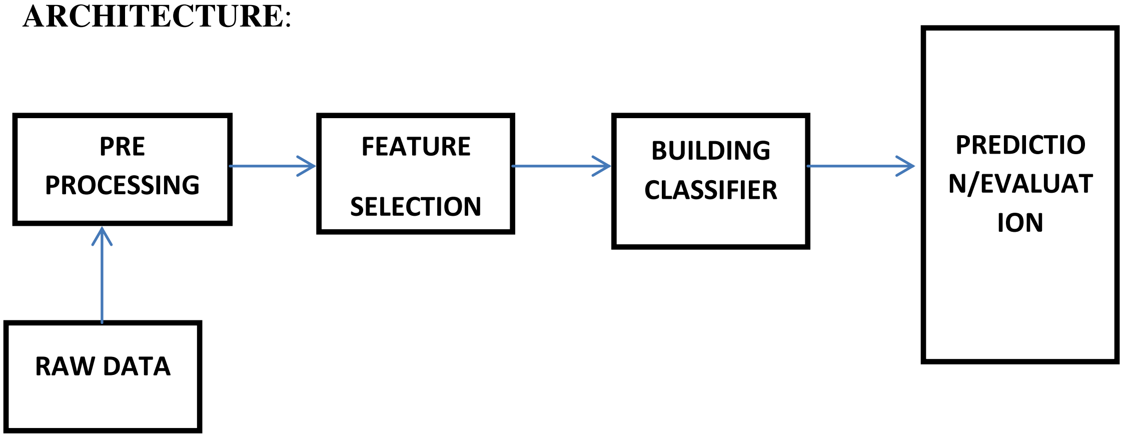


Fig .1 proposed Architecture

5. Data set used

The details of Summary and Description of calls that are noted in the tickets of the Health provider's ticketing system are the data that is used for classification. These records have already been classified by the call center professionals. Hence this is a Supervised learning Task.

The categories and Sub-categories distribution is as in the following table. I have created a new categorical variable that catches the essence of both categories and sub_categories I to the new derived_categories variable.

categories	**sub_categories**	**derived_categories** (22 UNIQUE CATEGORIES)	**No of Samples**
APPOINTMENTS	NEW APPOINTMENT	NEW APPOINTMENT	10478
	QUERY ON CURRENT APPOINTMENT	QUERY ON CURRENT APPOINTMENT	657
	RESCHEDULING	RESCHEDULING	1626
	RUNNING LATE TO APPOINTMENT	RUNNING LATE TO APPOINTMENT	694
	CANCELLATION	APPOINTMENTS CANCELLATION	417
ASK_A_DOCTOR	SYMPTOMS	SYMPTOMS	1201
	MEDICATION RELATED	MEDICATION RELATED	10598
JUNK	JUNK	JUNK	21
LAB	SHARING OF LAB RECORDS (FAX, E-MAIL, ETC.)	SHARING OF LAB RECORDS (FAX, E-MAIL, ETC.)	1425
	LAB RESULTS	LAB RESULTS	2650
	CANCELLATION	LAB CANCELLATION	246
MISCELLANEOUS	SHARING OF HEALTH RECORDS (FAX, E-MAIL, ETC.)	SHARING OF HEALTH RECORDS (FAX, E-MAIL, ETC.)	3550
	CHANGE OF	CHANGE OF	953

	PROVIDER	PROVIDER	
	OTHERS	OTHERS	7356
	QUERIES FROM INSURANCE FIRM	QUERIES FROM INSURANCE FIRM	107
	CHANGE OF HOSPITAL	CHANGE OF HOSPITAL	149
	CHANGE OF PHARMACY	CHANGE OF PHARMACY	55
PRESCRIPTION	REFILL	REFILL	9819
	PROVIDER	PROVIDER	1972
	QUERIES FROM PHARMACY	QUERIES FROM PHARMACY	1722
	PRIOR AUTHORIZATION	PRIOR AUTHORIZATION	1226
	FOLLOW UP ON PREVIOUS REQUEST	FOLLOW UP ON PREVIOUS REQUEST	357

6. DATA PREPROCESSING:

LANGUAGE USED: R language is used for Data Pre-processing and building of the Algorithm.

6.1 INITIAL INSIGHTS:

- The data is provided in the form of a CSV file by the company, where the DATA column of the table has RTF components which have to be removed before applying bag of words.
- After the removal of RTF components in the 'DATA' Column of the table the text from the 'SUMMARY' and 'DATA' columns is concatenated in to 'DATA' column and the 'SUMMARY' column is dropped.
- Other columns in the table such as 'field', 'previous appointment' and 'ID' have no significance on the classification task and are removed. The essence of the 'categories' and 'sub_categories' is captured in a single column 'Derived Category'. Hence the Classification in to 'Derived Category' is done based on the column 'DATA'
- There is a lot of class imbalance in the Data and the number of Data Samples is 57280 records which should be used for training and testing across 22 derived categories

7. STEPS FOR GETTING FURTHER INSIGHTS:

Summary of the call and description of the calls written by various staff members with no standard text guidelines. This along with the possibility of spelling mistakes deems the problem not straight forward.

In order to get insights in to the word distribution for each category, to know the ‘keywords ‘ in the text which decide the category and to correct the most probable spelling mistakes, the following steps are performed:

1. Combining all the ‘DATA” values of rows of the same category in to a corpus. Thus total 22 different corpuses are formed each for one category.
2. Constructing document term matrix for each category.(After performing stemming and stop word removal)
3. The no of samples for each category is noted
4. From the frequency distribution of terms of each category set of words(S1) with h frequency >0.2*no of samples of that category are noted. These are the words with highest frequency
5. From the frequency distribution of terms of each category set of words (S2) with frequency <=50 and>10 are noted down.
6. Seeing if the words in set S2 are misspelled words in S1 and noting them down.

8. PUTTING INSIGHTS IN TO ACTION:

The insights after applying the above-mentioned steps are:

- cx'd | cxl | cx | ca | cancell are the different abbreviations/misspelling for the word ’cancel’ which is a key word in derived_categories like ‘LAB CANCELLATIONS’ and ‘APPOINTMENTS CANCELLATION’
- pharm | pharmcy | phrmacy are the different abbreviations/misspelling for the word pharmacy which is a key word in derived_categories like ‘QUERIES FROM PHARMACY’, ‘CHANGE OF PHARMACY’ etc.
- pa | prior-authorisation | prior-authorization | pre-auth are forms of prior authorization found widely in ‘PRIOR AUTHORIZATION’ and ‘QUERIES FROM INSURANCE FIRM’.
- resh | rs | resch | resched and schedlue | scheduel | scheduld | sched | shed | sch are forms of reschedule and schedule which are key words in derived_categories like ‘NEW APPOINTMENT’, ’LAB CANCELLATION’, ‘APPOINTMENTS CANCELLATION’ etc.
- There are 17245 and 9671 unique words in the categories ‘MEDICATION RELATED and ‘REFILL’. Most of these words are the names of medicines that are prescribed. All these medicine names can be replaced with a word ‘medtype’ to decrease the number of dimensions.

antibiotic | amoxycillin | ibrufen | aspirin | oxycodoneacetaminophen | hydrocodone | hydrocodoneacetaminophen | imitrex | oxycodone | percocet | tylenol | promethazine | diclofenac | meclizine | phenergan | cyclobenzaprine | hydrobromide | levetiracetam |

nortriptyline | gabapentin | amitriptyline | trazodone | tecfidera | propranolol | prozac | lorazepam | dexamethason | guanfacine | methadone | metoclopramid | morphine | chlorzoxazone | lisinopril | nortriptyline | amphetamine | etodolac | oxcarbazepine | rizatriptan | baclofen | diphenhydramine | divalproex | ketorolac | lamotrigine | methocarbamol | metoprolol | carbamazepin | focalin | indomethacin | lisdexamfetamine | melatonin | naproxen | lovastatin | amantadine | amiodarone | ropinirole | antidepressant | clopidogrel | ethosuximide | tizanidine | bisacodyl | buspirone | dihydrochloride | dulcola | duloxetin | eletriptan | hydrochlorothiazide | levothyroxine | naratriptan | zolmitriptan | vicodin | lemtrada | acetazolamide | topomax | memantine | indomethacine | rivastigmine | lisinopril | lithium | meperidin|phentermine | olanzapine | rozerem | sinimet | sumavel | amitriptylin | provigil | risperidone | zanaflex | nortryptiline | focalin | tizanidine | benzoate | escitalopram | quetiapine | treximet | meloxicam | dexmethylphenide | eletriptan | coumadin | cyproheptadine | daytrana | metoclopramide | ketorolac | xarelto | pregabalin | imitrex | benadryl | medrol | percocet | ergocalciferol | dilantin | gabapentin | tegretol | tysabri | tramadol | indomethacin | neurontin | carbidopalevodopa | carbidopa | levodopa | ciprofloxacin | cholecalciferol | acylcarnitine | coumadin | methylprednisolone | topamax | nortriptylin | baclofen | clonazepam | tecfidera | ativan | diazepam | effexor | meclizine | toradol | valium | aggrenox | carbamazepine | tegretol | clonazepam | copaxone | zolpidem | belsomra | benadryl | triptans | triptan | triptane | carbatrol | trazadone | odansetron | phenergan | phenobarbitone | advil | dilaudid | flexeril | oxycontin | memantine | butalbital-acetaminophen-caffeine | rizatriptan | dextroamphetamine | duloxetine | zyprexa | benzoate | betaseron | percocet | copaxone | demerol | hcl | tylenol | imitrex | risperidone | ativan | lisdexamfetamine | eletriptan | topomax | zarontin | ergocalciferol | indomethacin | pravastatin | tizandine | warfarin | clopidogrel | aggrenox | amrix | buspar | tapentadolzolmitriptan | phenobarbital | phenobarbitone | quetiapine | tegretol | folic | cyclobenzaprine | butorphanol | rivastigmine | trihexyphenidyl | dilantin | dexmethylphenidate | effexor | robaxin | dihydrochloride | escitalopram | ondansetron | pregabalin | rituxan | ketoralac | bupropion | naproxen | entacapone | galantamine | ropinirole | selegiline | temazepam | brintellix | clobazam | fludrocortisone | focalin | pyridostigmine | tromethamine | amantadine | frovatriptan | phenytoin | vicoprofen | chlordiazepoxideamitriptylin | chlordiazepoxide-amitriptyline | mysoline | lorzone | acetazolamide | aripiprazole

- Replacing all time realted words to one word ‘timetype’
  monday | tuesday | wednesday | thursday | friday | saturday | mon | tue | wed | thu | fri | sat | january | february | march | april | may | june | july | august | september | october | november | december | afternoon | morning | today | yesterday | tomorrow "," timetype
- Replacing all labtest related terms to one word ‘labtype’
  radiologist | neurolab | anesthetist | hematology | echo | anesthesiologist | angiography | cisternogram | culture | diagnosis | mri | labwork | study | eeg | scan | psg | ncs | test | urine | screen | machine | doppler | cholesterol | cardio | methylmalonic | mrilab | urineanalysis | urine | plasma | count | scatter | cortisol | myelogram | mylogram
- Replacing different specialisations of doctors to one word ’doctype’

dsurgeon | dentist | neurologist | cardiologist | pedeatrician | orthopedic | neurology | neurosurgeon | ophthalmologist | dermatologist | oncology | pschychiatry | dermatology | hematologist | neuropsychiatrist | rheumatologist |dermatologist | cardiologist | allergist | audiologist | urologist | psychiatrist | podiatrist | plastic surgeon | physiologist | pediatrician | ent | oncologist | obstetrician | neurologist | neonatologist | immunologist | gynecologist | endocrinologist

- Removing Punctuations
- Removing Digits
- Stripping white space
- Building the corpus from 'DATA' column of the 57,280 records
- Building a document Term matrix, with the following options:

-tf
-tf-idf
-binary weighing

This is a Multiclass Classification problem on Text classification with total Classification Categories as 23. There is also a class imbalance in the training data. We will be exploring different Bag of Words Schemes and different machine learning algorithms.

9. THE FOLLOWING METHODS WILL BE EXPLORED:

APPROACH 1:

1. SVM on TF-IDF weighted Document Term Matrix.
2. KNN, Decision Trees and Random Forest on Tf-IDF weighted Document Term Matrix.
3. KNN, Decision Trees and Random Forest on Tf weighted Document Term Matrix.
4. KNN, Decision Trees and Random Forest on Binary weighted Document Term Matrix.
5. KNN, Decision Trees and Random Forest on Tf-IDF(normalized)weighted Document Term Matrix.
6. KNN, Decision Trees and Random Forest on log weighted Tf-IDF Document Term Matrix.

APPROACH 2:

7. Combining (column binding) binary tf and tf-idf in to a combined Document Term Matrix and applying KNN, Decision Trees and Document Term Matrix.

APPROACH 3:

8. Selecting the best features of the Weighted binary tf matrix using decision tree and appending the best features with tf-idf to build a random forest model.

APPROACH 4:

9. Combining tf, tf-idf and binary weighted tf and applying SVD to reduce the dimensions. Building a Random Tree model on the resultant features.

APPROACH 5:

10. Adding a new feature length which has the length of each document (number of words in each document) and appending this to Tf-Idf matrix and best features of Weighted binary tf and running a Random Forest.

10. RESULTS

APPROACH 1:

SVM on Tf-Idf Document Term Matrix gave an accuracy of **58**.

The approach followed was class based C with class weights given in the ration of 1/n1:1/n2:.....1/n21:1/n22, where n1,n2,n3,.....n22 are the no of samples of each respective class and a linear kernel was used.

The accuracy achieved is not cross-validated accuracy and SVM was taking about 10 hours for training.

onTf-IDF weighted Document Term Matrix

KNN: **59.5627**

Decision Trees: **44.245**

Random Forests: **65.68**

On Tf-IDF weighted (normalised) Document Term Matrix

KNN: **60.54**

Decision Trees: **37.38**

Random Forests: **65.69**

On log tf weighted Document Term Matrix

KNN: **60.24736**

Decision Trees: **44.24569**

Random Forests: **65.37**

On tf weighted Document Term Matrix

KNN: **59.73**

Decision Trees: **44.24**

Random Forests: **65.47**

On weighted binary Document Term Matrix

KNN: **58.64**

Decision Trees: **37.03**

Random Forests: **64.37**

APPROACH 2:

Usingcombinedbinary tf and tf-idf

KNN: **59.75**

Decision Trees: **44.24**

Random Forests: **65.50**

APPROACH 3:

Selecting the best features of the Weighted binary tf matrix using decision tree and appending the best features with tf-idf to build a random forest model.

Accuracy: **69.415**

APPROACH 4:

Combining tf, tf-idf and binary weighted tf and applying SVD to reduce the dimensions. Building a Random Tree model on the resultant features.

Accuracy: **66**

All of the above approaches(except SVM) were done on 3-fold cross-validation. Train and Test are acquired by stratified split of ratio 75:25. All the above models have been run on the same split.

11. CONCLUSION:

In this paper we have implemented different machine learning algorithms like decision tree, Random forest and KNN with various techniques TF-IDF, Normalized Term frequency and accuracy of the models are calculated. These models can able to classify the tickets described in the text into different classes that will made the job of departments for effectively serving the patients.

REFERENCES:

1. Jiang, M.; Liang, Y.; Feng, X.; Fan, X.; Pei, Z.; Xue, Y.; Guan, R. Text classification based on deep beliefnetwork and softmaxregression.NeuralComput. Appl.2018,29, 61–70.2.
2. Kowsari, K.; Brown, D.E.; Heidarysafa, M.; JafariMeimandi, K.; Gerber, M.S.; Barnes, L.E. HDLTex:Hierarchical Deep Learning for Text Classification.Machine Learning and Applications (ICMLA).In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications(ICMLA), Cancun, Mexico, 18–21 December 2017.3.

3. McCallum, A.; Nigam, K. A comparison of event models for naive bayes text classification. In Proceedings ofthe AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, USA, 26–27 July 1998; Volume 752,pp. 41–48.
4. Kowsari, K.; Heidarysafa, M.; Brown, D.E.; JafariMeimandi, K.; Barnes, L.E. RMDL: Random MultimodelDeep Learning for Classification. In Proceedings of the 2018 International Conference on Information Systemand Data Mining, Lakeland, FL, USA, 9–11 April 2018; doi:10.1145/3206098.3206111.
5. Heidarysafa, M.; Kowsari, K.; Brown, D.E.; JafariMeimandi, K.; Barnes, L.E. An Improvementof Data Classification Using Random Multimodel Deep Learning (RMDL).IJMLC2018,8, 298–310,doi:10.18178/ijmlc.2018.8.4.703.6.Lai, S.;
6. Xu, L.; Liu, K.; Zhao, J. Recurrent Convolutional Neural Networks for Text Classification.In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA,25–30 January 2015; Volume 333, pp. 2267–2273.
7. Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms.InMining Text Data; Springer:Berlin/Heidelberg, Germany, 2012; pp. 163–222.
8. Aggarwal, C.C.; Zhai, C.X.Mining Text Data; Springer: Berlin/Heidelberg, Germany, 2012.
9. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval.Inf. Process. Manag.1988,24, 513–523.
10. Goldberg, Y.; Levy, O. Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embeddingmethod.arXiv2014, arXiv:1402.3722.
11. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedingsof the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar,25–29 October 2014; Volume 14, pp. 1532–1543.
12. Mamitsuka, N.A.H. Query learning strategies using boosting and bagging.InMachine Learning: Proceedingsof the Fifteenth International Conference (ICML'98); Morgan Kaufmann Pub.: Burlington, MA, USA, 1998;Volume 1
13. Chen, W.; Xie, X.; Wang, J.; Pradhan, B.; Hong, H.; Bui, D.T.; Duan, Z.; Ma, J. A comparative study of logisticmodel tree, random forest, and classification and regression tree models for spatial prediction of landslidesusceptibility.Catena2017,151, 147–160
14. Larson, R.R. Introduction to information retrieval.J. Am. Soc. Inf. Sci. Technol.2010,61, 852–853.
15. Xu, B.; Guo, X.; Ye, Y.; Cheng, J. An Improved Random Forest Classifier for Text Categorization.JCP2012,7, 2913–2920.