Data Quality Improvement for Medical Concept Normalization

Krishna Vikas Meduri
Data Science
University of North Texas
Denton, Texas, USA
krishnavikasmeduri@my.unt.edu

Rashmi Jakkani
Data Science
University of North Texas
Denton, Texas, USA
rashmijakkani@my.unt.edu

Harith Mote
Data Science
University of North Texas
Denton, Texas, USA
harithmote@my.unt.edu

Prudhvi Narayan Korupolu
Data Science
University of North Texas
Denton, Texas, USA
PrudhviNarayanKorupolu@my.unt.e

***Abstract*—Automatically recognizing medical concepts mentioned in social media texts are used to enhance the people health quality in a community. However, the discrepancy between the language used in social media (such as twitter and Facebook) and the medical concepts have been challenging. Evaluated the approaches using three different datasets, where social media texts are extracted from twitter or Facebook messages. Poor data quality has also direct impact on the machine learning systems. This first, exposed from a dataset that has hidden problem that is used to built a machine learning concepts. Due to the poor quality of the dataset and the defective validation process the results of the experiment showed the great performance. The results also demonstrated that the evaluation of data quality is useful for guiding the quality improvement of machine learning. Therefore, proposed a data quality evaluation framework that includes the quality and their corresponding approaches.**

The code and the data for this project are available on GitHub at: https://github.com/krishnavikas-7/INFO-5731/blob/main/Term_Final_Project.ipynb

Keywords—Data Quality, Medical concept normalization, Exploratory Data Analysis, Model implementation, Agreement rate

I. INTRODUCTION

The data present in social media can be used to improve the understanding of patience experience in healthcare such as the spread of infectious diseases and side-effects of drugs. However, the social media text and medical concepts text have discrepancy between them and that has become very challenging. Particularly, the frequent use of informal language and abbreviation forms, as well as the social media messages has taken into account by effective information extraction systems.

Deep learning as the most foremost vital breakthrough in machine learning history that has drawn extraordinary consideration from academics as well as business. Be as that it may, missing of high quality preparing information gets to be a major risk to the utilization of deep learning. Approaches, such as crowdsourcing web information or exchanging information from other spaces have been proposed for upgrading preparing information. By the by, this information seems present commotions, such as invalid information and name clamors. Although some carefully outlined profound learning models are vigorous to massive name commotions, the computing rule the show of "garbage in garbage out" is still pertinent to profound learning.

Rajpurkar detailed a profound learning framework called ChexNet that was developed for diagnosing pneumonia illness based on chest X-ray pictures. They claimed that the ChexNet surpassed normal radiologist execution on pneumonia discovery on both sensitivity and specificity. Many radiologist and machine learning experts doubted the result due to the improper dataset. The problem is due to the lacking approval of the measurable methods with genuine information within the fMRI thinks about. Numerous tests have shown that information with destitute quality may contrarily affect the performance of profound learning altogether. Other experiments have illustrated that way better quality of training data might move forward the execution of profound learning. Hence, an efficient assessment of the quality of the dataset is critical for building a high-quality machine learning framework. The evaluation result would offer rules for information enhancement and framework execution enhancement.

One of the purposes of using this is to improve the understanding the impact of data quality on the performance of machine learning systems. Informative quality is a multi-dimensional concept on subjective and quantitative properties of information, and the definition changes beneath diverse settings.

In this article, we characterize data quality as a estimation of data for fitting the reason of building a machine learning framework. We begin with explored an overclaimed execution enhancement of a machine learning framework. The issue was due to the destitute quality of the datasets for building the framework and the risky approval handle. At last, we presented a system for evaluating the information quality to guarantee the quality necessities of datasets for building a high-quality machine learning framework. It was built on profound learning models, which require a huge sum of preparing information. The datasets displayed in this article appear deficiently and there's a

large parcel of cover between the preparing and the test information.

Data Quality is the measure of how well suited a data set is to serve its specific purpose. Measures of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness and timeliness.

- Improved data quality improves the performance and accuracy of the models that have been deployed. You can have more confidence in your performance if you have more High-Quality data. Better data can help to decrease hazards while also improving consistency of outcomes.
- As a result, we present a data quality evaluation framework that comprises the quality criteria as well as the evaluation methodologies that match to them. Machine learning may make use of the data validation procedure, performance improvement method, and data quality evaluation framework.
- The effect of data quality on the performance of data-driven machine learning systems. Data quality is a multifaceted notion that encompasses both qualitative and quantitative data qualities, and its meaning changes depending on the situation.

This project, which aims to improve the health of individuals in a community, recognizes medical ideas stated in social media communications automatically, allowing for a variety of applications. The information gleaned from social media can be used to better understand patient experiences in healthcare, such as the transmission of infectious diseases and medicine side effects. Effective information extraction systems must account for the widespread use of informal language, non-standard grammar and abbreviation forms, as well as social media posts. The task of medical concept normalization for social media text is comparable to that of mapping a variable length social media message to a medical concept in some external coding system. We go beyond string matching in this project. Using deep neural networks, we suggest learning and exploiting the semantic similarity between texts from social media postings and medical concepts.

Social media message	Description of corresponding medical concept
lose my appetite	Loss of appetite
i don't hunger or thirst	Loss of Appetite
hungry	Hunger
moon face and 30 lbs in 6 weeks	Weight Gain
gained 7 lbs	Weight Gain
lose the 10 lbs	Body Weight Decreased
feeling dizzy ...	Dizziness
head spinning a little	Dizziness
terrible headache!!	Headache

The above figure describes the aim of the project precisely, in the first instance the social media message depicts that loss of appetite which is related to the medical concept loss of appetite.

In the same way the medical problem was also conveyed as in the social media "I don't have hunger or thirst" which is also relates to the medical concept "loss of appetite".

In this way we have tried to annotate the social media tweets by relating both tweets and concept definitions.

This process allows us to assess many unanswered social media medical problems.

The normalization was built on a neural-network-based text classification to learn the mapping between social media messages and medical concepts. They conducted the tests on the three datasets. The comes out of the tests detailed in appeared the models with a dialect show, which was pretrained on Google news, achieved the most excellent execution among all the three datasets regarding exactness separately. Be that as it may, a precise approval is required to affirm the issue. We created an approach for expelling the clamors in the dataset and conducted a precise examination. We found that normalization execution diminished when the sum of noises diminished within the datasets.

II. Literature Review

Data collection, cleaning and accuracy is highly important when it comes to modelling a dataset for analysis or prediction. 80% of the process modelling lies in data cleaning and imputation. By far, there are many approaches that include meticulous research and proper data representation techniques.

One of these is the research on "Normalising medical concepts in Social Media Texts by learning semantic representation". In this paper the author discussed on how using convolution neural network and recurrent neural network has benefited in improving the overall dataset performance by learning the semantic meaning of the dataset before sending it to the classifier. The difference when compared to previous works is that this model captures the dependency between the terms and represents the social media text message into a lower dimensional vector before mapping it to the medical concepts.

In Another approach, "Data Evaluation and Enhancement for Quality Improvement of Machine Learning" it is discussed that using transfer learning can be used to address the data quality issue and performance of the machine learning model. It is identified that the existing models were good enough because of the misleading dataset and poor validation process. Transfer learning based strategy is used to get the relation between data quality and the efficiency of the model chosen to work with social media texts. Meanwhile, In the proposed approach we explored building the data quality using the labelled data from various clinical knowledge sources for medical concept normalization from user generated social media texts.

We constructed a set of experiments to study the research problems using the aforementioned datasets, pretrained

language models, and deep learning models. These experiments' settings are summarized. The following is a description of some of these experiments. Experiments to look at the medical concept normalization system's overclaimed performance issue. The outcomes of the experiment will help us establish a data-driven approach for better verifying machine learning systems and understanding how data quality influences their performance. The fine-tuning approach was used in the trials.

To evaluate the effectiveness of fine-tuning to performance improvement, ULM Fit was first fine-tuned with dataset and ask a Patient for medical concept normalization. Experiments with extra datasets to fine-tune the medical concept normalization systems and discover how the target dataset affects machine learning performance. The performance of the ULMFit-based medical concept normalization models that were fine-tuned with datasets Cadec, Pubmed, Health news, Big tweet, TwADR-L, and Ask a Patient was compared. TwADR-L and Ask a Patient datasets were used to train and evaluate the medical concept normalization classifier.

III. METHODS

- Data annotation plays a key role in medical concept normalization. It is a process of labeling the data which notably valuable in supervised machine learning (ML), where the system processes, understands, and learns from input patterns to produce desired outputs. Analysing linguistic nuances for modelling discourse is critical. So here each concept is labelled by everyone. If the Definition of the Concepts and Cleaned Tweets are Relevant, then it is Annotated as "1" if the Definition and Cleaned Tweets are Not Relevant it is Annotated as "0". To perform the annotation, we implemented majority voting strategy to gain adequate results. Majority voting is a popular and robust strategy to aggregate different opinions in learning from crowds, where each worker labels examples according to their own criteria.

A	B	C	Majority
1	0	1	1
0	0	1	0

Table :1

In the above table, we can see that two individuals labelled as '1'(relevant), so the majority is '1'. For the next row two individuals annotated a s'0'(irrelevant),so the majority is considered as irrelevant. With the help of majority voting strategy the quality of the dataset will be improved because majority voting includes knowledge and opinions of different individuals will be taken into consideration and only one accurate annotation will be voted as the input parameter to the perform machine learning operations to gain desired outputs.
After performing the majority voting we can check the agreement rate (Kappa value). Agreement rate (Kappa Value) is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured.

Kappa value interpretation:
<0 No agreement
0 — .20 Slight
.21 — .40 Fair
.41 — .60 Moderate
.61 — .80 Substantial
.81–1.0 Perfect

The above interpretation illustrates the kappa scores for the majority voting to each individual labelled data. Here .81-1.0 denotes that there is high relevancy in that data and vice versa.

IV. DATA COLLECTION AND CLEANING

Measures of data quality are based on data quality characteristics such as accuracy, completeness, consistency, validity, uniqueness, and timeliness. To ensure this purpose the first step is collection of relevant valid data. We have divided the dataset into 3 categories namely TwADR-S, TwADR-L, AskPatient. In this data, the message is labelled with a single category name amongst the chosen categories. TwADR-S is the dataset provided by Collier and Limsopatham, which has 201 twitter tags and texts and their corresponding concepts. The total number of targeted concepts in the dataset are 58 in which the mean medical concept can be mapped by 3.47 queries.
TwADR-L is the dataset provided by the same in which three months of twitter data is collected using Twitter Streaming API. This data consists of 4 drug profiles with 1436 phrases that can be mapped to approximately 2300 medical concepts.
The final category is Ask Patient which is less ambiguous and less informal than the other two due to its clarity. It consists of 8662 twitter phases that can be mapped to 1036 medical concepts.

After all the collection process our final dataset consists of medical concepts, definitions of medical concepts and the cleaned social media tweets.
By verifying both concept definition and cleaned tweets we were able to annotate the data which helps us to implement further process.

- Statistical analysis:

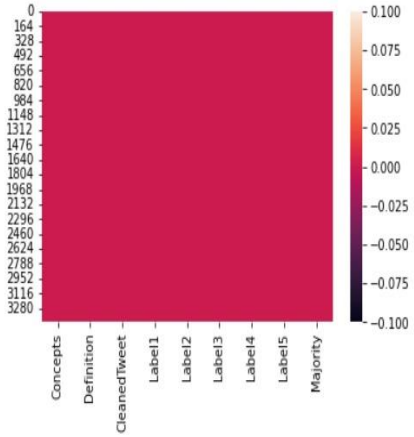
```
[21] temp = df.groupby('Majority').count()['CleanedTweet'].reset_index().sort_values(by='CleanedTweet',ascending=False)
     temp.style.background_gradient(cmap='Purples')
```

	Majority	CleanedTweet
1	1	2075
0	0	1358

The above table indicates the number of relevance(1) and irrelevance(0) data annotations in the data.

```
sns.heatmap(df.isnull())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f87512da610>
```



The above heat map illustrates that there are no null values in the dataset.

```
[15] df.describe()
```

	Label1	Label2	Label3	Label4	Label5	Majority
count	3433.00000	3433.000000	3433.000000	3433.000000	3433.000000	3433.000000
mean	0.60501	0.507719	0.510632	0.503641	0.493737	0.604428
std	0.48892	0.500013	0.499960	0.500060	0.500034	0.489045
min	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.00000	1.000000	1.000000	1.000000	0.000000	1.000000
75%	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000
max	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000

```
[16] df.shape
     (3433, 9)
```

pd.describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values

- Data Pre-processing:

The process of converting raw data into a comprehensible format is known as data preparation. We can't work with raw data; thus this is a key stage in data mining. Before using machine learning or data mining methods, make sure the data is of good quality.

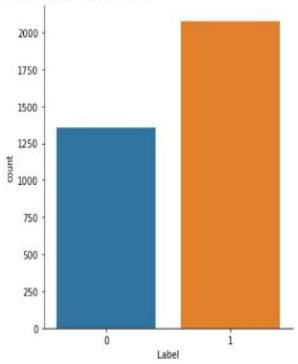
```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
df["Majority"] = le.fit_transform(df["Majority"])
le.classes_
output: array([0, 1])

from scipy import stats
stats.mode(df.Majority)
output:ModeResult(mode=array([1]),
count=array([2075]))

from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])

import nltk  #cleaning
df['CleanedTweet'] = df['CleanedTweet'].str.replace('[^\w\s]','')  #removal of punctuation
df['CleanedTweet'] = df['CleanedTweet'].apply(lambda x: " ".join(x.lower() for x in x.split()))#lower case
from nltk.corpus import stopwords #removal of stopwords
nltk.download('stopwords')
stop = stopwords.words('english')

df['CleanedTweet'] = df['CleanedTweet'].apply(lambda x: " ".join(x for x in x.split() if x not in stop))
words = []  #building document-term matrix
from nltk.tokenize import RegexpTokenizer
from gensim import corpora,models
tokenizer = RegexpTokenizer(r'\w+')
for x in pd.Series(df['CleanedTweet']):
  a = tokenizer.tokenize(x)
  words.append(a)
dictionary = corpora.Dictionary(words)
corpus = [dictionary.doc2bow(word) for word in words]

import gensim  #LDA model
model = gensim.models.ldamodel.LdaModel(corpus, num_topics = 10, id2word = dictionary,passes = 20)

print(model.print_topics(num_topics = 10, num_words = 5
```

V. Experiment and Data analysis

Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigation on data to discover patterns, to spot anomalies, to test hypothesis with the help of Statistics and Graphical Representation.

- To comprehend the Label column's categorical variables To establish the number of Labels, a cat plot visualization was used. Catplot displays the frequencies of one, two, or three category variables, and it is a new feature in seaborn.

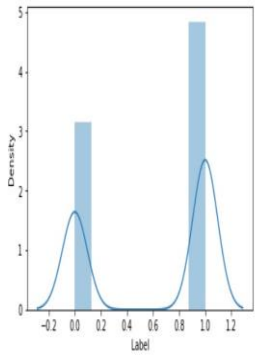
```
In [ ]: sns.catplot(x='Label',data=df,kind='count')
        df.Label.value_counts()

Out[ ]: 1    2077
        0    1356
        Name: Label, dtype: int64
```



- Distplot represents the overall distribution of continuous data variables. The Distplot depicts the data by a histogram and a line in combination to it.

Concepts

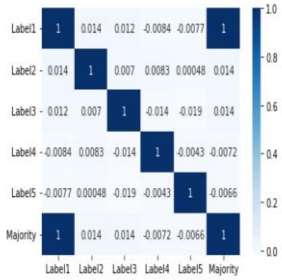
```
In [ ]: sns.distplot(df['Label'])

        D:\New folder\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a depr
        n. Please adapt your code to use either `displot` (a figure-level function with similar flexibilit
        s).
          warnings.warn(msg, FutureWarning)
Out[ ]: <AxesSubplot:xlabel='Label', ylabel='Density'>
```



- Correlation is a statistical term that expresses how closely two variables are related in a linear fashion (meaning they change together at a constant rate). It's a typical way of describing simple relationships without stating a cause-and-effect relationship.

Correlation in the data

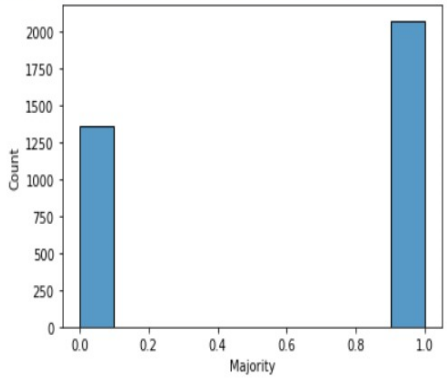
```
In [109...  plt.figure(figsize=(6,4))
            sns.heatmap(df.corr(),cmap='Blues',annot=True)

Out[109... <matplotlib.axes._subplots.AxesSubplot at 0x7fa52532bf50>
```



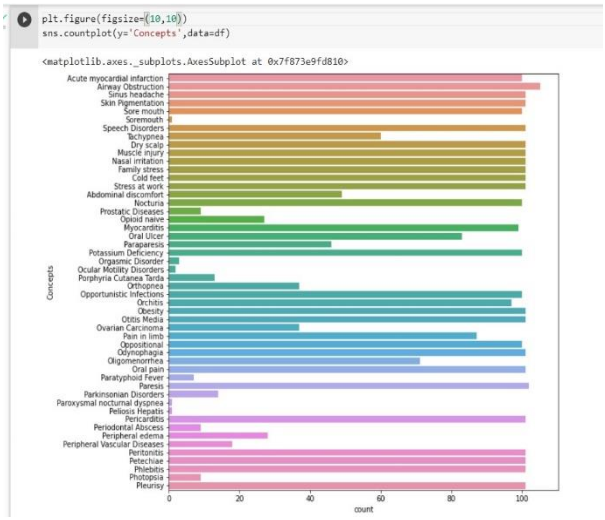
Dark shades represents positive correlation while lighter shades represents negative correlation.

- A histogram is a graph that allows you to identify and display the underlying frequency distribution (shape) of continuous data. This enables data to be examined for its underlying distribution (e.g., normal distribution), outliers, skewness, and other factors.

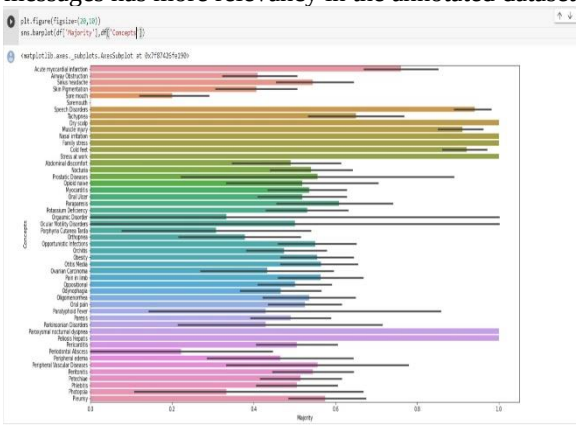
```
✓ [26] sns.histplot(df.Majority,bins=10)
0s
       <matplotlib.axes._subplots.AxesSubplot at 0x7f8747b3f250>
```



- The count plot below represents most frequent concepts in the dataset.



- The barplot below represents the which concept social messages has more relevancy in the annotated dataset



VI. Results and Discussion

- Feature extraction

  The term frequency — inverse document frequency, or TF-IDF, is a scoring measure commonly employed in information retrieval (IR) or summarization. The TF-IDF is used to determine how important a phrase is in a particular document. Here the TF-IDF is performed on the cleaned tweets to get the weightage of each word in the cleaned tweets.

- Training and Evaluation of the Models

  1.Naive Bayes:

  Naive Bayes Algorithm is used to predict the probability of different classes based on various attributes.

  The Naive Bayes Algorithm is majorly used in Text Classifications and to deal with problems having multiple Classes.

  2.Support Vector Machine:

  Support Vector Machine are linear classifiers which are based on the margin maximization principle. They perform structural risk minimization, which improves the complexity of the classifier with the aim of achieving excellent generalization performance.

  3.Random Forest:

  A random forest is a machine learning technique for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers.

  Many decision trees make up a random forest algorithm. Bagging or bootstrap aggregation are used to train the 'forest' formed by the random forest method. Bagging is a meta-algorithm that increases the accuracy of machine learning methods by grouping them together.

  4.Decision tree:

  Decision Tree is a Supervised Machine Learning Algorithm that makes judgments based on a set of rules, like how people do.

  A Machine Learning classification algorithm can be thought of as being created to make judgements. You might claim that the model predicts the class of a new, never-before-seen input, but the algorithm has to select which class to assign behind the scenes. Some classification algorithms, such as Naive Bayes, are probabilistic, but there is also a rule-based approach.

  5.KNN:

  The supervised machine learning algorithm k-nearest neighbours (KNN) is a simple, easy-to-implement technique that may be used to address both classification and regression issues. Both classification and regression predicting issues can be solved with KNN. However, in the industry, it is more commonly employed in categorization difficulties. We look at three key criteria while evaluating any technique:

  1. The output is simple to interpret.

  2. Time to calculate

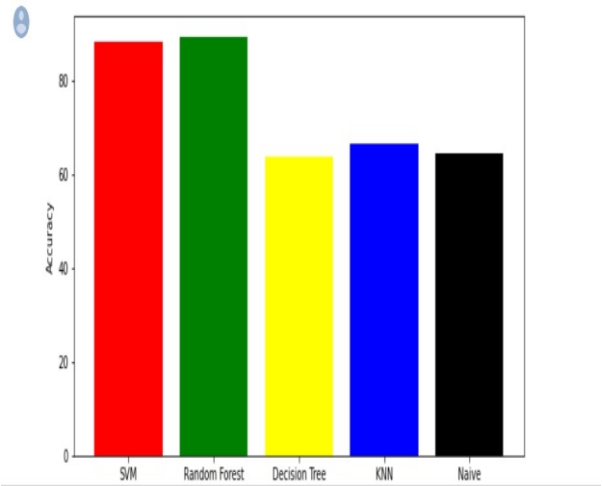
  3. Predictive Ability

  The supervised machine learning algorithm k-nearest neighbours (KNN) is a simple, easy-to-implement technique that may be used to address both classification and regression issues.

- Accuracy scores:

- Accuracy visualizations:

```
import matplotlib.pyplot as plt
fig = plt.figure(figsize = (10,5))
algo=['SVM','Random Forest','Decision Tree','KNN','Naive']
acc=[]
acc.append(svm_accuracy*100)
acc.append(rf_accuracy*100)
acc.append(dt_accuracy*100)
acc.append(knn_accuracy*100)
acc.append(naive_accuracy*100)
plt.bar(algo,acc,color=['red','green','yellow','blue','black'])
plt.xlabel('Algorithms')
plt.ylabel('Accuracy')
plt.show()
```



Models	Accuracy
Naïve Bayes	81.60%
Svm	87.5%
Decision tree	87.9%
Random forest	88.8%
KNN	67.40%

VII. Conclusion

The Main Aim of this project is to Annotate the data for the medical concepts and the social media messages relating to the medical concepts. We Annotate the data by verifying the relevancy between the definition of the concept and the social media message if the tweet and the definition of the message is relevant we annotate it as 1(relevant) or else we annotate it as 0(not irrelevant) by performing this annotation we can be able to improve the performance of the machine learning by giving the accurate input parameters and to obtain desired outputs. We've implemented the majority Vote strategy to give the best outcome where the annotations of everyone in the group are taken into consideration and check their relevancy to the concept and take a vote among the group for to choose the better annotated data and perform it in the machine learning models. Also, by checking the agreement rate (Kappa Values) we can get the interrater reliability between each annotated column the kappa value is calculated by the observed agreement and the agreement by chance. After applying all the strategies which will best suit for further processing, we've pre-processed the data by cleaning and extracting the stop words and nonstandard grammar and also performed Exploratory and Statistical Data Analysis for the better understanding of the dataset.

Cross validation score:

Models	Cross validation score
Naïve Bayes	62%
SVM	63.9%
Decision tree	62.7%
Random forest	63.3%
KNN	62.1%

The Above Table depicts that SVM has Highest Cross Validation Score.

Once the EDA process is done The Machine Learning models are implemented. Models implemented are Naïve-Bayes, KNN, SVM, Decision Tree, Random Forest. The Random Forest Model got the highest Accuracy Score where as SVM & Decision followed by second and third best accuracy scores and Naïve bayes and KNN got least accuracy among the others. This Result made us understand that Random Forest Model is better suited to the dataset.

VIII. Limitations and future work

The Difficulties and Limitations we've faced during the project is annotating the medical concepts and some of the tweets are not accurate enough to check the relevancy between the concept definition and the cleaned tweet. Our Future Work is to improve more quality and accuracy to the dataset by comparing other related datasets which could help to get a better assessment of the patient's experience and it can also be helpful to prevent and control false medical news in the social media and it can also be able to help to control the viral spreads across the internet.

.

IX. Author Contributions

Krishna Vikas Meduri - team lead, writing for Literature review, Methodology, Data collection and cleaning, Data annotation, general editing, Model implementation, Leading team meetings.

Harith Mote -Methodology, Experiment and data analysis, data annotation, Model Implementation, Results and discussions participating in team meetings

Rashmi Jakkani – Introduction, Abstract, Data annotation, Model Implementation, Methodology, Experiment and Data analysis, Data cleaning, Scheduling and participating in team meetings.

Prudhvi Narayan Korupolu – Data annotation

References:

[1] Y. Bengio, "Learning deep architectures for AI," Found. Trends Mach.
Learn., vol. 2, no. 1, pp. 1–127, 2009.

[2] D. Silver et al. "A general reinforcement learning algorithm that masters
chess, shogi, and Go through self-play," Science, vol. 362, no. 6419,
pp. 1140–1144, 2018.

[3] J. Krause et al., "The unreasonable effectiveness of noisy data for finegrained recognition," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 301–320.

[4] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning:
Transfer learning from unlabeled data," in Proc. 24th Int. Conf. Mach.
Learn., 2007, pp. 759–766.

[5] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust
to massive label noise," CoRR, vol. abs/1705.10694, 2017: Accessed: 19
Apr. 2021. [Online]. Available: http://arxiv.org/abs/1705.10694

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy
disparities in commercial gender classification," in Proc. Conf. Fairness,
Accountability, Transparency, 2018, pp. 77–91.

[7] P. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on
chest X-rays with deep learning," CoRR, vol. abs/1711.05225, 2017. Accessed: 19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1711.05225

[8] L. Oakden-Rayner, "Exploring the ChestXray14 dataset: Problems," 2017.
Accessed: Jun. 19, 2019. [Online]. Available: https://lukeoakdenrayner.
wordpress.com/2017/12/18/the-chestxray14-dataset-problems

[9] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fMRI
inferences for spatial extent have inflated false-positive rates," Proc. Nat.
Acad. Sci. USA, vol. 113, no. 28, pp. 7900–7905, 2016.

[10] E. Giannoulatou, S.-H. Park, D. Humphreys, and J. Ho, "Verification and
validation of bioinformatics software without a gold standard: A case study
of BWA and Bowtie," BMC Bioinformat., vol. 15(Suppl16), 2014, Art. no.
S15.

[11] J. Zhang et al., "Analysis of cellular objects through diffraction images
acquired by flow cytometry," Opt. Exp., vol. 21, no. 21, pp. 24819–24828,
2013.

[12] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable
effectiveness of data in deep learning era," inProc. IEEE Int. Conf. Comput.
Vis., 2017, pp. 843–852.

[13] C. Su and D. Huang, "Hybrid recommender system based on deep learning
model," Int. J. Performability Eng., vol. 16, no. 1, pp. 118–129, 2020.

[14] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the meaningfulness of 'big data quality'," Data Sci. Eng., vol. 1, no. 1, pp. 6–20,
2016.
[15] H. Chen, G. Cao, J. Chen, and J. Ding, "A practical framework for
evaluating the quality of knowledge graph," in Proc. China Conf. Knowl.
Graph Semantic Comput., 2019, pp. 111–122.
[16] J. R. Evans and W. M. Lindsay, The Management and Control of Quality,
vol. 5. Cincinnati, OH, USA: South-Western, 2002, pp. 115–128.
[17] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means
to data consumers," J. Manage. Inf. Syst., vol. 12, no. 4, pp. 5–33, 1996.
[18] N. Limsopatham and N. Collier, "Normalising medical concepts in social
media texts by learning semantic representation," in Proc. 54th Annu.
Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers), 2016, pp. 1014–
1023.
[19] K. Lee, S. A. Hasan, O. Farri, A. Choudhary, and A. Agrawal, "Medical
concept normalization for online user-generated texts," in Proc. IEEE Int.
Conf. Healthcare Informat., 2017, pp. 462–469.
[20] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep
transfer learning," in Proc. Int. Conf. Artif. Neural Netw., 2018, pp. 270–
279.
[21] ILSVRC2012, "ImageNet large scale visual recognition challenge," 2012.
Accessed: Jun. 20, 2019. [Online]. Available: http://www.image-net.org/
challenges/LSVRC/2012/results.html
[22] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," Commun. ACM, vol. 39, no. 11, pp. 86–95, 1996.
[23] O. Bodenreider, "The Unified Medical Language System (UMLS): Integrating biomedical terminology," Nucleic Acids Res., vol. 32, no. suppl_1,
pp. D 267–D270, 2004.
[24] D. Britz, "Implementing a CNN for text classification in TensorFlow,"
Accessed: Jan. 10, 2020. [Online]. Available: http://www.wildml.com/
2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/
[25] "Recurrent neural networks and LSTM tutorial in Python and
TensorFlow." 2017. Accessed: Jan. 10, 2020. [Online]. Available:
https://adventuresinmachinelearning.com/recurrent-neural-networkslstm-tutorial-tensorflow/
[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed
representations of words and phrases and their compositionality," in Proc.
Adv. Neural Inf. Process. Syst., 2013, pp. 3111–3119.
[27] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of
fine-tuning deep learning models for plant disease identification," Comput.
Electron. Agriculture, vol. 161, pp. 272–279, 2019.
[28] J. Howard and S. Ruder, "Universal language model fine-tuning for text
classification,"CoRR, vol. abs/1801.06146, 2018. Accessed: 19 Apr. 2021.
[Online]. Available: http://arxiv.org/abs/1801.06146
[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-trainingof
deep bidirectional transformers for language understanding," CoRR, vol.
abs/1810.04805, 2018. Accessed 19 Apr. 2021. [Online]. Available: http:
//arxiv.org/abs/1810.04805
[30] "The TwADR-L dataset." 2016. Accessed: Feb. 10, 2020. [Online]. Available: https://zenodo.org/record/55013#.XK-1zOtKh24
[31] "Healthcare Twitter analysis." 2016. Accessed: Feb. 10, 2020. [Online].
Available: https://github.com/grfiv/healthcare_twitter_analysis
[32] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text
classification?" inProc. China Nat. Conf. Chin. Comput. Linguistics, 2019,
pp. 194–206.
[33] "The WikiText long term dependency language modeling dataset—
WikiText-103." 2020. Accessed: Jan. 10, 2020. [Online]. Available:
https://blog.einstein.ai/the-wikitext-long-term-dependency-languagemodeling-dataset/
[34] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for
word representation," in Proc. Conf. Empirical Methods Natural Lang.
Process., 2014, pp. 1532–1543.
[35] S. Faltl, M. Schimpke, and C. Hackober, "ULMFiT: State-of-the-art
in text analysis." 2019. Accessed: Jan. 10, 2020. [Online]. Available: https://humboldt-wi.github.io/blog/research/information_systems_
1819/group4_ulmfit/
[36] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing
LSTM language models," CoRR, vol. abs/1708.02182, 2017. Accessed:

19 Apr. 2021. [Online]. Available: http://arxiv.org/abs/1708.02182
[37] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM Comput. Surv., vol. 41, no. 3, pp. 1–52, 2009.