Novel Approach on Topic Modelling paradigms for Recognition of health-related topics in social platform through Distributed file system

Abstract- social media is one of the most prominent features of the Internet and has a vast amount of influence in today's world. All the latest information regarding every topic is available on social media, particularly Twitter where discourse on every major trending topic takes place and is shown to the user hence it is utilized by many individuals all over the world to keep themselves updated with the latest news occurring everywhere in the world. In Current time due to the pandemic, Twitter also played a key role in relaying updated and accurate information regarding many topics related to health care. Millions of people everyday share their ideas and opinions regarding health-related concepts. To analyze these many tweets which contain billions of users is a challenging task and so to understand the key features or topics discussed by the user's artificial intelligence has a branch called Natural language processing also called as NLP which can help in analyzing texts from a huge amount of dataset. Topic modeling, which is a part of NLP helps in analyzing healthcare data extracted from twitter which would provide a description of current trending healthcare topics and development of most significant tweets and their content.

Keywords: Healthcare, Artificial Intelligence, Topic modeling

1. Introduction

Twitter consists of a vast number of resources and data which can be utilized to implement machine learning algorithms and artificial intelligence for different use case scenarios like healthcare systems. The rapid growth in the amount of technology used by the health system could always use this processed data for more advanced research systems. Healthcare systems that were developed earlier used past data to understand the current requirements that very well could be outdated in the current scenario. So, it is more effective to use the updated current data extracted from Twitter. Social media is generally faster than traditional media and the news is unadulterated so we can see the people's opinion around a particular topic without any biases. Most of the Twitter data is in the form of text so to analyze this enormous data which consists of tweets sent by billions of users NLP algorithms are more suitable. Topic modeling algorithms which come under nlp can be utilized to analyze this textual data and understand the key features or the current trending topics discussed by people relating to healthcare. Initially the data from Twitter is entirely unstructured and needs to be preprocessed for the system to be more efficient. Various pre-processing techniques are applied to the textual data and the resultant output is then subjected to feature extraction techniques to identify the most important recurring topics and their related words. The resultant output of these algorithms is a vector which consists of different topics and words that are related to these topics. One of the significant aspects regarding information on social media is the spread of incorrect information, in healthcare it could be false symptoms, incorrect diagnosis etc. This spread of false information can then be visualized by

checking how frequent the related keywords are occurring for a particular topic. The final output of all the algorithms is then used to create a word cloud that gives a concise report on the most used word in the processed tweets.

2. Literature Survey

[1]. Kannan's research is an analysis of 571 documents up to 2022. The research contains topics like covid 19, operations and supply chain management in the health industry. This research used NNMF, LDA, LSI, the algorithms that are used in the project. Pre-processing operations like stop word elimination is done. The outcome is a data cloud that consists of all the keywords that are used for further analysis.

[2]. Harris's research is based on patient satisfaction. Patients are asked for a review of their experience and 33000 reviews are collected. Topic modeling approach is used to reveal negative reviews and complaints. In this research LDA algorithm is used for topic modeling. Then Naive Bayes algorithm is used for topic classification. Their research mainly focused on the LDA algorithm and helped us get a thorough understanding of the LDA algorithm and learn how efficient the algorithm is in creating topic models.

[3]. Giorgi's paper provides information on content relating to specific LDA topic modeling techniques which can then be utilized to recognize different domains of covid specific discourse that can then be utilized to track concerns and views in current social environment. This paper also suggests that model obtained topics are more consistent when compared to the standard LDA topics and another benefit is that they give new useful features that helps in predicting new Covid-19 related outcome.

[4]. This paper gives us a comparative study of LDA and LSA topic models. The paper performs the similarity computation among the given subjects by using the LDA and LSA topic models. It also shows us the computational costs of building the LSA and LDA matrices.

[5]. The steps that are needed before applying LDA to textual data include appropriate pre-processing, adequate selection of model parameters, evaluation of model's reliability, and process of validly interpreting results. This paper aims to provide more information and accessibility to researchers in LDA topic modeling and to create a hands-on technique to apply topic modeling.

3. Methodology

The Multiple steps involved in the project methodology are represented in the below diagram. They include Data extraction, pre-processing, feature extraction, applying topics modeling algorithms and finally visualization.

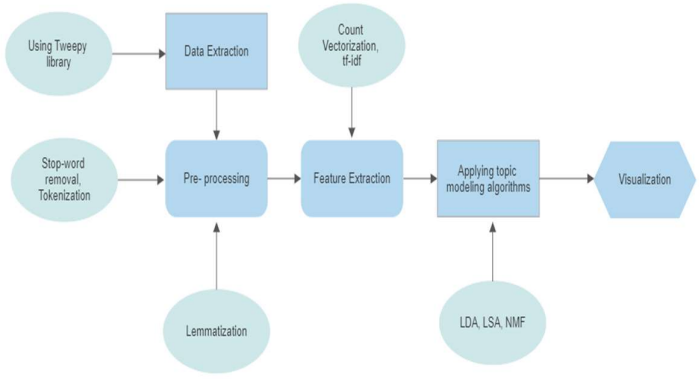

Fig. 1 Workflow Diagram

The process starts by gathering data from Twitter about various health diseases that are prevalent around the world. To gain access to the dataset, a Twitter developer account is first needed. This can be obtained by following the given steps.

1. Log in to your Twitter account and verify your email address or if your Twitter account doesn't exist sign up to twitter.com.
2. Sign up to **developer.twitter.com** after logging in and enter your developer account name, purpose of use and the country.
3. Read the developer agreement, click accept button and submit.
4. log in to your email and verify your developer account.
5. A new app can be created via the Developer Portal.

3.1 Dataset

To extract the dataset from Twitter, a new project needs to be created in the developer portal from which tokens are generated which can be used for extracting the dataset. In python, the Tweepy library is utilized to import dataset from Twitter. Tweepy is an open-source library that helps in accessing twitter API with python. This library includes classes that have Twitter's models and API endpoints. It can be used to do a variety of tasks such as decoding and encoding data, OAuth authentication, HTTP requests, Streams and Rate limits.

3.2 Pre-Processing the data.

Once the data has been acquired from the Twitter API, it needs to be pre-processed. Since the raw dataset is in the form of texts. There are certain pre-processing techniques which can be applied to clean the dataset.

3.2.1 Stop-Word Removal: Stop word removal is a pre-processing technique used in nlp applications. Just deleting the words that appear in every document in the corpus is the notion. Pronouns and articles are typically categorized as stop words. In some NLP tasks, such as information retrieval and classification, these words have no value, therefore getting rid of them increases the algorithm's effectiveness.

3.2.2 Tokenization: One of the first steps in the NLP pipeline is Tokenization. This has important implications for the remaining pipeline. Tokenization tears down data that is unstructured and other natural language texts into different chunks of information that can be viewed as discrete elements. Tokenization can be utilized to separate different words or sub words, sentences, and characters.

Tokenization is done in the following sequence:

- Tokenize the text into sentences.
- Tokenize sentence into words.
- Tokenize regular expression sentences.
- The output can be used as input to other algorithms.

3.2.3 Lemmatization: Lemmatization is a common text pre-processing technique used in natural language processing (NLP). Lemmatization usually refers to lexical and morphological analysis of a word to get things right, usually with the goal of removing only inflections and returning the dictionary version of the word generally known as the lemma. Lemmatization usually simply folds the various inflections of the lemma. Group words with similar meaning into one word.

After applying the following pre-processing techniques, all the unnecessary data is removed like retweets, links, user id's, symbols, numbers, hashtags etc. All the uppercase words are converted to lowercase for further simplification. The dataset is transformed from unstructured data to a stream of textual data containing string that holds valuable data.

3.3 Feature Extraction

Feature Extraction techniques can be utilized to reduce many different features in a dataset by producing new features from the currently existing features and then replacing them. The new reduced features contain crucial information of the original features that were removed. There are 2 proposed features extraction techniques used in the project.

3.3.1 Count Vectorization: Computers don't have the ability to understand letters and words. So, when working with text data, it must be represented numerically so that it can be understood by machines. The text is parsed to do some predictive modeling to remove words. This is a process called tokenization. These words should be encoded as numbers to be used as input for machine learning algorithms. A Count vectorizer is a way of converting text into numbers that the system can understand. Count vectorizer makes it easy to use text data directly in deep learning models such as machine learning and text classification. This can be used to remove stop words. A stop word is a word that without changing the meaning of the sentence can be safely excluded from the sentence. It works by checking each word count in the document. CountVectorizer in the python's sci-kit library can be utilized to implement this.

3.3.2 Term Frequency (TF) - Inverse document frequency (IDF): It is one of the major algorithms used for converting text into meaningful numeric representations and is used to adapt machine algorithms for prediction. Machine learning algorithms are realized through mathematical elements such as statistics, algebra and calculus. The problem with natural languages is that the data is in the form of raw text, so we need to convert the text to vectors. tf-idf vectorizer considers the overall weight of the words

in the document. It combines two concepts: term frequency (TF) and inverse document frequency (IDF).

- Term frequency signifies the total number of times a particular text appears within the document.
- IDF represents occurrence of a particular term in the whole corpus of the documents. So, the main thing to notice is that it's common to all the documents. Document frequency is the number of documents containing a particular term. By giving higher weight to rare terms and lower weight to common terms, stop words can be removed very effectively. So, if the word is more frequent and appears in multiple different documents, its idf value will reach to 0, alternatively if the word is rare its idf value will reach to 1.

3.4 Topic Modeling Algorithms

Topics modeling is generally utilized for unsupervised information and possesses a clear difference from text classification and clustering tasks, while text classification and clustering is used to make information retrieval easy and make clusters of documents, In topic modeling the text is distributed and there are multiple topics. In topic modeling, generally clusters of three types of words are made co-occurring words, histogram of topic wise words and distribution of words. Algorithms that are utilized for topic modeling are Latent Dirichlet Allocation, Correlated topic modeling, Latent semantic analysis and Probabilistic latent semantic analysis. Topic mining techniques are extensively utilized for text mining tasks. This approach is mainly utilized for long format content and is less effective for short text format. It is mainly used in machine learning to find thematic relations in a mainly huge collection of documents with textual data.

There are many applications of topic modeling with supervised, unsupervised, and semi-supervised approaches being updated and invented to apply in machine learning, text classification, text mining, recommendation engines, and information retrieval. Playing a major role in Information retrieval and in Natural language processing. Topics modeling is primarily performed on document repositories with textual information or data. Information retrieval in the application involves representation of documents, the ranking system, queries, and the framework. Topic modeling is also used in providing meaningful textual classification in the databases of genomics which usually have large amounts of textual content. In genomics the search engines used to apply topic modeling to collate and present relevant information to the used. The applications of topic modeling are very simple, but the vast array of methodologies used to sort and represent the information is very crucial and important.

Algorithms utilized in topic modeling:

3.4.1 Latent Semantic Analysis (LSA): LSA is one of the major topic modeling algorithms which can be used to extract topics from a set of documents by converting their text into document topic matrix and word topic. The steps taken in LSA are relatively simple. We convert the text data into a document term matrix then subsequently implement a truncated singular value decomposition (SVD) and finally encode the words with the topics that are extracted. The LSA model calculates the frequency of words that occur within documents and across texts and assumes that similar documents have approximately the same distribution of word frequencies for particular words. Syntactic and semantic information is ignored, and each text document is treated as a set of words. The Standard method for calculating frequency of words is known as tf-idf.

3.4.2 Latent Dirichlet Analysis (LDA): The foundation of both LDA and LSA is the same: distributional hypothesis, which holds that documents discuss a variety of issues

for which a statistical distribution may be calculated, and statistical mixing hypothesis, which holds that documents discuss a range of topics. The core idea of LDA is to map each document in our corpus to a set of topics that include every word in the document. LDA assigns topics to a collection of words in order to map the documents. This results from the hypothesis that documents generally consist of collections of words and those collections dictate topics. LDA assumes that distribution of vast topics in a set of documents and distribution of a set of words are Dirichlet distributions.

There are two terms alpha and beta which refer to two criteria that regulate the similarity of documents and topics. Each document will have fewer topics if the alpha value is low, whereas a high value will have the reverse effect. Another third hyperparameter also exists which must be set to implement LDA which will compute the number of different topics the algorithm will be able to detect.

The result of the algorithm consists of a vector that will cover predominantly every topic of the document. In which the first part will display the first topic and the rest will be similar [0.5, 0.3, 0.1, ...]. When these vectors are compared, they will give you the insights into the topical characteristics of your corpus.

3.4.3 Non-Negative Matrix Factorization (NMF): NMF is an unsupervised approach that is necessary to rebuild the original data by projecting data into lower dimensional areas. A collection of multivariate analysis and linear algebra procedures where a matrix P is factored into two matrices Q and R, all of which have non-negative members. The generated matrices are simpler to inspect thanks to this characteristic. Non-negativity is a property of the data being evaluated in applications like processing audio spectrograms or muscle activity. The problem is commonly approximated numerically because it is typically not fully solvable.

4. Results

4.1 Count Vectorization

Fig.2. Result after applying Count Vectorization

4.2 LSA

```
LSA Output
Topic 0:
['monkey', 'happened', 'covid', 'like', 'vaccine', 'thing', 'rabies', 'know', 'people', 'shit']
Topic 1:
['rabies', 'like', 'vaccine', 'dont', 'covid', 'dengue', 'people', 'cancer', 'think', 'shots']
Topic 2:
['covid', 'cancer', 'dengue', 'diabetes', 'breast', 'obesity', 'people', 'cases', 'year', 'like']
Topic 3:
['cancer', 'diabetes', 'breast', 'awareness', 'monkey', 'obesity', 'month', 'risk', 'insulin', 'cure']
Topic 4:
['diabetes', 'insulin', 'million', 'americans', 'type', 'year', 'past', 'study', 'rationed', 'obesity']
Topic 5:
['dengue', 'fever', 'cases', 'vaccine', 'symptoms', 'takeda', 'approval', 'test', 'mosquito', 'patients']
Topic 6:
['obesity', 'health', 'causes', 'food', 'like', 'alzheimer', 'disease', 'people', 'know', 'real']
Topic 7:
['alzheimer', 'health', 'caregivers', 'ones', 'struggle', 'loved', 'disease', 'real', 'natures', 'dementia']
Topic 8:
['year', 'insulin', 'million', 'study', 'americans', 'past', 'rationed', 'finds', 'ration', 'alzheimer']
Topic 9:
['like', 'people', 'looks', 'alzheimers', 'look', 'dont', 'disease', 'yeah', 'insulin', 'study']
```

Fig.3. Results after applying LSA.

4.3 LDA

```
LDA Output
Topic 0:
['like', 'people', 'rabies', 'free', 'said', 'thats', 'love', 'data', 'shot', 'medical']
Topic 1:
['alzheimer', 'early', 'treatment', 'caregivers', 'ones', 'struggle', 'loved', 'happened', 'brain', 'alzheimers']
Topic 2:
['alzheimers', 'health', 'disease', 'alzheimer', 'dementia', 'real', 'natures', 'people', 'research', 'read']
Topic 3:
['know', 'awareness', 'dont', 'month', 'need', 'time', 'hospital', 'dengue', 'help', 'week']
Topic 4:
['monkey', 'covid', 'make', 'really', 'better', 'person', 'death', 'hard', 'want', 'didnt']
Topic 5:
['dengue', 'cases', 'care', 'fever', 'good', 'thing', 'october', 'news', 'days', 'country']
Topic 6:
['rabies', 'dengue', 'covid', 'vaccine', 'think', 'today', 'test', 'virus', 'work', 'live']
Topic 7:
['cancer', 'obesity', 'breast', 'risk', 'women', 'help', 'research', 'great', 'causes', 'diagnosed']
Topic 8:
['look', 'patient', 'blood', 'government', 'body', 'researchers', 'kill', 'silo', 'university', 'pharma']
Topic 9:
['diabetes', 'year', 'study', 'insulin', 'million', 'type', 'fighting', 'high', 'americans', 'past']
```

Fig.4. Results after applying LDA.

4.4 NMF

```
NMF Output
Topic 0:
['monkey', 'happened', 'vaccine', 'thing', 'shit', 'disappeared', 'know', 'getting', 'heard', 'think']
Topic 1:
['rabies', 'vaccine', 'dont', 'shots', 'bite', 'vaccines', 'anti', 'shot', 'think', 'airborne']
Topic 2:
['covid', 'vaccine', 'good', 'long', 'know', 'died', 'time', 'cases', 'strain', 'booster']
Topic 3:
['cancer', 'breast', 'awareness', 'month', 'cure', 'risk', 'october', 'know', 'patients', 'diagnosed']
Topic 4:
['diabetes', 'type', 'people', 'high', 'adults', 'sugar', 'risk', 'mellitus', 'clinical', 'trial']
Topic 5:
['dengue', 'fever', 'cases', 'vaccine', 'symptoms', 'takeda', 'test', 'approval', 'mosquito', 'patients']
Topic 6:
['obesity', 'causes', 'food', 'health', 'america', 'unhealthy', 'problem', 'know', 'proud', 'healthy']
Topic 7:
['alzheimer', 'ones', 'caregivers', 'struggle', 'loved', 'health', 'real', 'natures', 'disease', 'dementia']
Topic 8:
['year', 'insulin', 'million', 'americans', 'study', 'past', 'rationed', 'diabetes', 'finds', 'ration']
Topic 9:
['like', 'looks', 'people', 'yeah', 'gone', 'look', 'sounds', 'alzheimers', 'cure', 'dont']
```

Fig. 5 Results after applying NMF.

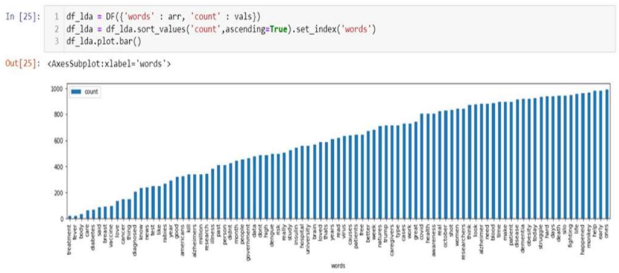

Fig. 6. Bar graph of words vs frequency

The above graph shows the popularity of each word where each word is represented in X-axis and their frequency is represented in Y-axis.

5. Conclusion

From the above project we have successfully analyzed 4000 unique tweets related to 8 different diseases and have extracted key concepts discussed world-wide. 3 topic modeling algorithms LDA, LSA, NMF is used to gain the key concepts from the thoughts sharing platform called as twitter. Each algorithm is used to produce 10 different topics having 10 key words each. An overall of 100 unique topics are to be generated. The topics generated by LDA, LSA, NMF are each combined and then used to generate a graph that gives a count of how many topics are generated and which topic appears how many times in the given set of documents. LDA has given 91 topics, LSA produced 56 topics, NMF generated 85 topics. By making a comparative study we can confirm that LDA is the best algorithm used. LDA will be finally used to generate a word cloud giving a precise and concrete representation of the occurrences of topics.

6. Future Work

The current taken dataset is a mix of 8 different diseases and the dataset is manually extracted. This slows the project as we cannot get the information about any other related disease. Developing an interactive chatbot makes it easier for the user to understand the concept of the project quickly and put it for good use. The accuracy and efficiency of the algorithms of the project can be much more improved with a feature extraction concept called PCA [Principal component analysis]. The project can be much more improved by focusing on one single disease which gives a perfect output about the key topics and words used in the discussion forum. An application that gets information of a particular country's health status can be developed for further accuracy.

References

1. Ali, I., Kannan, D. Mapping research on healthcare operations and supply chain management: a topic modeling-based literature review. Ann Oper Res 315, 29–55 (2022)
2. Doing-Harris K, Mowery DL, Daniels C, Chapman WW, Conway M. Understanding patient satisfaction with received healthcare services: A natural language processing approach. AMIA Annu Symp

3. Zamani M, Schwartz HA, Eichstaedt J, Guntuku SC, Ganesan AV, Clouston S, Giorgi S. Understanding Weekly COVID-19 Concerns through Dynamic Content-Specific LDA Topic Modeling. Proc Conf Empir Methods Nat Lang Process. 2020 Nov; 2020:193-198.
4. Comparing LDA and LSA Topic Models for Content-Based Movie Recommendation Systems: Sonia Bergamaschi & Laura Po
5. Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri & S. Adam (2018) Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology, Communication Methods and Measures.
6. Liang Jie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). Association for Computing Machinery, New York, NY, USA, 80–88.
7. Resnik, Philip, et al. "Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter." Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality.
8. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter: Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, Tingshao Zhu