Characterizing diabetes, diet, exercise, and obesity comments on Twitter

Pankaj Kumar Magar - 222IT018
Information Technology
National Institute of Technology,
Karnataka, Surathkal, India 575025
pankajkumarmagar19@gmail.com

Vedant Parwal - 222IT034
Information Technology
National Institute of Technology,
Karnataka, Surathkal, India 575025
vedant2000parwal@gmail.com

Dr. Sowmya Kamath
Information Technology
National Institute of Technology,
Karnataka, Surathkal, India 575025
sowmyakamath@nitk.edu.im

***Abstract*— The prevalence of diabetes, obesity, and related comorbidities continues to increase worldwide, leading to significant public health concerns. Social media platforms like Twitter provide a unique opportunity to explore public perceptions and attitudes toward these health issues. In this study, we aimed to characterize diabetes, diet, exercise, and obesity-related comments on Twitter.This study aims to examine the attributes of the opinions held by the general population concerning diabetes, diet, exercise, and obesity (DDEO) as conveyed on the social media platform Twitter.We collected a dataset of tweets containing keywords related to diabetes, diet, exercise, and obesity from January 2020 to June 2021 using Twitter's API. We applied natural language processing techniques to preprocess and analyze the tweets, including sentiment analysis, topic modeling, and network analysis.Overall, our study provides insights into the public perceptions and attitudes towards diabetes, diet, exercise, and obesity on Twitter. These findings could be used to inform public health campaigns and interventions aimed at promoting healthy lifestyle choices and preventing chronic diseases.**

***Keywords:* Health, Diabetes, Diet, Obesity, Topic model, Text mining, Twitter**

I. INTRODUCTION

Diabetes, obesity, and other related health issues are among the most prevalent chronic diseases in the world. These health conditions pose a significant public health concern as they lead to severe complications and increased mortality rates. With the rise of social media platforms, public attitudes and perceptions towards these health issues can be studied more effectively. In recent years, Twitter has become a prominent platform for sharing health-related information, making it an ideal source for studying public opinions and discussions about diabetes, diet, exercise, and obesity (DDEO).

Almost 1.9 billion persons are now regarded to be overweight, and over 650 million adults are now considered to be obese, according to the World Health Organization.Based on the Study of Obesity by the European Association, overweight and obese rank as the fifth highest risk factor for global deaths. Increased consumption of calories and inadequate energy expenditure together with the accumulation of extra body weight result in weight gain and an increased risk of developing diabetes. Obesity can be successfully reduced by the use of modifiable lifestyle habits including diet and exercise. Numerous comorbidities, such as diabetes, are frequently present in people who are overweight or obese.

This study's goal is to categorise tweets that mention DDEO using a multi-faceted semantically and linguistic framework. In order to gain insight into the public's knowledge, attitudes, and opinions concerning DDEO, the study set out to gather Twitter data, identify interesting subjects, and analyze these topics.25 percent of the 92000 tweets that were retrieved referenced diabetes, 25 percent nutrition, 25 percent exercise, and 25 percent obesity. Previous research on Twitter has concentrated on finding recurring themes connected to a particular health condition that users have discussed. In contrast, this study adopts a fresh strategy by employing computational analysis to look at unstructured text data from Twitter that relates to health in order to gauge how the general public feels about four prevalent diseases: diabetes, diet, exercise, obesity (DDEO).

Table 1 DDEO queries

Health issue	Queries	Tweets	Percentage
Diabetes	diabetes OR #diabetes	23000	25 %
Diet	diet OR #diet	23000	25 %
Exercise	exercise OR #exercise	23000	25 %
Obesity	obesity OR #obesity	23000	25 %

II. LITERATURE SURVEY

Social media sites like Twitter have been a major source of information for studies connected to health in recent years. Several research have examined various health issues, such as obesity, diabetes, and exercise, using data from Twitter.

A study in (Edo-Osagie et al., 2020) found that the use of Twitter for public health research has grown rapidly in recent years, with a focus on four main areas: disease surveillance and outbreak detection, health behaviours and attitudes, health promotion and education, and health service utilization and delivery.The paper concludes that Twitter has the potential to be a valuable tool for public health research, particularly in the areas of disease surveillance and health promotion.

The short message lengths on Twitter make it difficult to fully utilise the traditional text mining tools that many researchers would like to use to interpret the posts. In the study (Hong and Davison, 2010) presents how the models can be trained on the dataset in order to overcome the issue of employing conventional topic models in microblogging contexts. We present alternative techniques to train a shared subject model while assessing how well they work using a set of painstakingly prepared trials for both quantitative and qualitative perspectives.

The clinical reports often contain a large amount of information that is difficult to extract and analyze using traditional methods, and that topic modelling can provide a more efficient and effective approach.In (Arnold and Speier, 2012) the author provides an overview of previous research on clinical text mining and topic modelling, highlighting the potential benefits of using topic modelling to extract key information from unstructured clinical reports.Large text corpora can be analysed using topic models to find word clusters that are semantically related. In this work, a topic model that supports specific patient timelines and is adapted to the clinical reporting context is presented.Findings demonstrate that the model can recognise patterns of clinical occurrences in a group of patients with brain tumours.

In the study (Dahal et al., 2019) the methodology used in the paper involves collecting a large dataset of tweets related to global climate change, preprocessing the data using natural language processing techniques, and then applying topic modelling and sentiment analysis algorithms to the preprocessed data. Specifically, the authors used Latent Dirichlet Allocation (LDA) for topic modelling and the Valence Aware Dictionary and Sentiment Reasoner (VADER) for sentiment analysis.This study's objective is to analyse a set of tweets using topic modelling, sentiment analysis, and volume analysis, and then compare the results over time and space.While volume analysis can be used to assess what argument is taking place in various locations or time periods, sentiment analysis is helpful in identifying the emotional state or viewpoint that is portrayed in the dataset.

III. METHODOLOGY

A. Data Collection

Using Twitter's Application Programming Interface (API), this phase gathered tweets. Diabetes, food, exercise, and obesity were chosen as the associated words and linked health topics inside the Twitter API. Both historical and real-time data gathering are offered through Twitter's APIs. Many pre-defined DDEO-related queries (Table 1) were utilised in this paper's real-time technique to randomly gather 10% of publicly accessible English tweets during a set period of time. The inquiries helped us gather almost 0.092 million linked tweets. The information was gathered using the twitter ids available on the website of the original author.

B. Data Cleaning

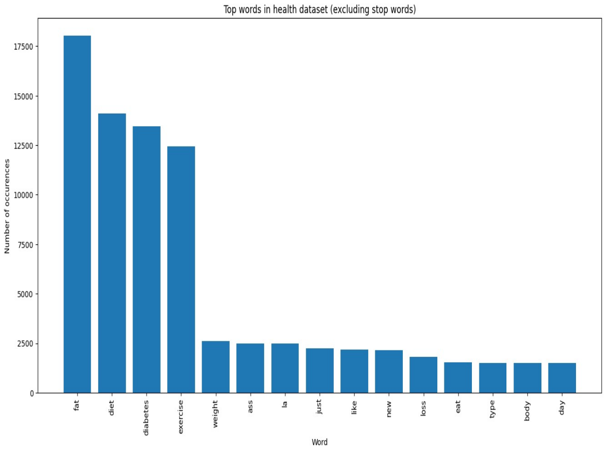


Fig. 1: Top Words in health dataset

To conduct topic modeling on the tweets, they must first be formatted in a way that is conducive to generating effective topics. This involves using a standardized list to eliminate stop words that lack semantic value for analysis (e.g. "the"), as well as removing hyperlinks and non-alphabetic characters like punctuation and numbers particularly from each tweet. The resulting message is then split by spaces to create a word list. To prevent the topics generated by the model from being control by the same top words, a list of common English words that lack intrinsic meaning (such as articles and conjunctions) as well as the keywords are removed. This process helps ensure that the resulting topics are meaningful. The following pertinent variables are included in each tweet's data: tweet id (the special ID assigned to the tweet), tweet (text), and date and time.

C. Topic Discovery

A type of statistical model termed topic discovery or modeling is used to recognize the abstract "themes" appearing in a group of documents. Similar to clustering on numerical data, topic modeling is unsupervised classification technique for documents that recognizes some natural groups of objects or subjects.Topic Modelling can help with:

- Discovering the hidden themes or topics in a document or corpus.
- Classifying the document into discovered themes or topics.

We utilized a topic modeling technique to pick out relevant topics from the pool of gathered tweets. This involved grouping together semantically similar phrases, like "diabetes," "cancer," and "influenza," into a broader theme such as "disease." In the health and medical sectors, topic modeling has a range of applications, such as predicting protein-protein connections by leveraging insights from existing literature. Here we are using four types of topic modelling

- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (TF-IDF)
- Latent Semantic Analysis (TF-IDF)

The most widely used topic model is Latent Dirichlet Allocation (LDA), which has been proved in research to be an efficient computer linguistics model for finding themes in a corpus. LDA makes the assumption that a corpus comprises themes that can each have a word assigned to it with varying degrees of membership.In huge document collections, latent topic information can be found using the unsupervised machine learning technique known as latent Dirichlet allocation. The method it employs is known as the "bag of words," and it views each page as a vector of word counts.

In LDA, a topic is essentially just a distribution function across the entire corpus of words. The quantity of subjects must be decided before executing LDA. The underlying premise is that each text in the corpus is a probabilistic blend of terms and topics. Using the premise that words that appear in the same document are more likely to pertain to the identical topic than words that do not, and that documents that include the same words are more likely to have the same themes than documents that do not, LDA analyses the word co-occurrences within texts. Steps for LDA :

- The processed data is lemmatized keeping only noun, adjective, verb using spacy.
- Using this lemmatized tokens, we form a dictionary and document term matrix.
- This dictionary and document term matrix is inserted into the mallet LDA model.

LSA stands for Latent Semantic Analysis, which is a computational technique used in NLP to discover and analyze the relationships between words and documents. LSA is a type of dimensionality reduction algorithm that identifies the underlying latent structure in a dataset.The primary goal of LSA is to transform the high-dimensional, sparse data representation of a collection of texts into a lower-dimensional, dense space that captures the most essential information about the relationships between the words and documents. This is accomplished by modelling every document then word as a series of vectors in a high-dimensional space, saving the most crucial data, and then applying linear algebra techniques to reduce the complexity of the space.

One of the advantages of LSA is that it can capture the underlying semantic relationships between words and documents, even when they are not explicitly stated.For example, LSA can identify that "cat" and "dog" are related words, even though they are not synonyms. LSA can also capture the meaning of words in context, which is important for many NLP tasks.The LSA algorithm works by constructing a matrix representing the frequency of occurrence of each word in each document. This matrix is known as the term-document matrix. The matrix is subsequently divided into three pieces via the singular value decomposition (SVD) process used by LSA, a matrix of left singular vectors, a matrix of distinct values, and a matrix of right singular vectors.

D. TF-IDF

Term Frequency-Inverse Document Frequency is referred to as TF-IDF. A corpus, or the collection of all texts, is a measure of statistics used to assess the significance of a term in a document.The total amount of times a term appears in a document is referred to as the term frequency (TF) of that phrase. The logarthmic ratio between the total amount of documents in the corpus and the number of documents containing a phrase is known as the inverse document frequency (IDF) of that term. Indicating that certain terms are more crucial for identifying a particular document, the IDF score is greater for terms that exist in fewer documents.The combination of the term frequency and the inverse document frequency specifies a term's TF-IDF score in a document. This score measures a term's prominence compared to other terms in the same text as well as other documents in the corpus. An important term for the document will get a high TF-IDF score, while a less important term will have a low TF-IDF score.

We can assign weights to the terms in the LDA model using the TF-IDF scores to combine LDA with TF-IDF. The name of this strategy is LDA-TF-IDF. The words in each text are given weights according to the LDA-TF-IDF scores, which are then employed in the LDA model to identify the subjects. We can make sure that the terms that are most important in each document possess a greater weight and are more probable to be assigned to its primary topics by applying TF-IDF weights.

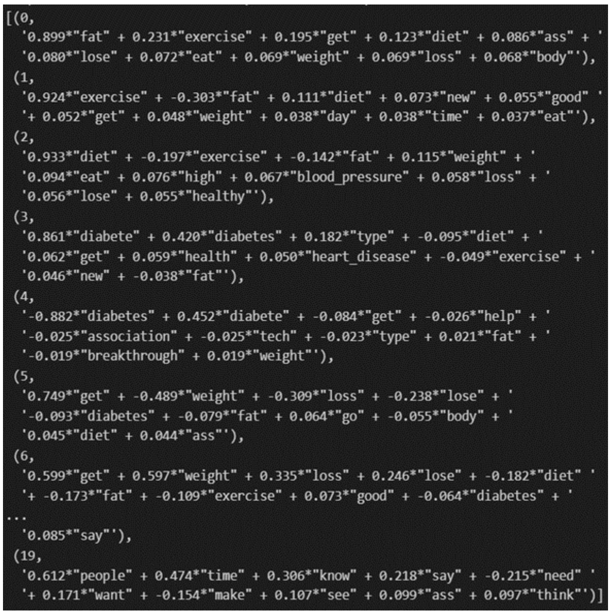
LSA and TF-IDF can be coupled to enhance the accuracy of both methods. TF-IDF-LSA is the name given to the union of LSA and TF-IDF. In this method, documents are initially represented as vector of weighed term frequencies using TF-IDF. LSA is then used on these vectors to extract the fundamental semantic structure and minimise the degree of dimensionality of the data. This resulting space in lower dimensions can subsequently be applied to information retrieval and document similarity. By reducing the noise and capturing the latent semantic connections between terms and documents, TF-IDF-LSA can enhance the performance of information retrieval and text classification tasks. It is a potent method for studying huge text corpora and extracting important data from them.

E. Bert Embedding

A "word embedding" is a mathematical depiction of a word in a continuous vector space used in Natural Language Processing (NLP). It is used to record the connections between the words and phrases and their semantic meaning. A particular kind of word embedding known as a "BERT embedding" (Bidirectional Encoder Representations from Transformers) creates representations of words or phrases using a neural network model known as a Transformer. A cutting-edge language model created by Google is called BERT. It may be tailored for a variety of NLP applications, like analysis of sentiment, query answering, and translation of languages. Unlike BoW and TF-IDF, BERT embeddings consider the surrounding words and sentences when generating the vector representation. BERT can capture complex relationships between words and phrases and provide better semantic understanding of the text.

IV. RESULTS AND ANALYSIS

Using LDA we found 222 topics linked to health with a coherence score of 0.5473. We also categorised topics based on the presence of DDEO terms. For instance, we classified a topic as diet-related if it contained the word "diet." Obesity, diabetes, nutrition, and exercise were hot topics, as was foreseen and required by the initial Tweeter API query (DDEO).The LDA algorithm discovered both DDEO and non-DDEO terms for each DDEO subject, which each included a number of linked subtopics (Table 2). In order of frequency, the most frequently discussed subtopics under "Diabetes" include type 2 diabetes, obesity, food, etc. Similar to this, Table 2 displays the subtopics discovered for each topic.The optimal topics obtained from the Topic modelling techniques which can also be used to calculate the correlation between the topics as shown in Fig 2.



```
[(0,
  '0.899*"fat" + 0.231*"exercise" + 0.195*"get" + 0.123*"diet" + 0.086*"ass" + '
  '0.080*"lose" + 0.072*"eat" + 0.069*"weight" + 0.069*"loss" + 0.068*"body"'),
 (1,
  '0.924*"exercise" + -0.303*"fat" + 0.111*"diet" + 0.073*"new" + 0.055*"good" '
  '+ 0.052*"get" + 0.048*"weight" + 0.038*"day" + 0.038*"time" + 0.037*"eat"'),
 (2,
  '0.933*"diet" + -0.197*"exercise" + -0.142*"fat" + 0.115*"weight" + '
  '0.094*"eat" + 0.076*"high" + 0.067*"blood_pressure" + 0.058*"loss" + '
  '0.056*"lose" + 0.055*"healthy"'),
 (3,
  '0.861*"diabete" + 0.420*"diabetes" + 0.182*"type" + -0.095*"diet" + '
  '0.062*"get" + 0.059*"health" + 0.050*"heart_disease" + -0.049*"exercise" + '
  '0.046*"new" + -0.038*"fat"'),
 (4,
  '-0.882*"diabetes" + 0.452*"diabete" + -0.084*"get" + -0.026*"help" + '
  '-0.025*"association" + -0.025*"tech" + -0.023*"type" + 0.021*"fat" + '
  '-0.019*"breakthrough" + 0.019*"weight"'),
 (5,
  '0.749*"get" + -0.489*"weight" + -0.309*"loss" + -0.238*"lose" + '
  '-0.093*"diabetes" + -0.079*"fat" + 0.064*"go" + -0.055*"body" + '
  '0.045*"diet" + 0.044*"ass"'),
 (6,
  '0.599*"get" + 0.597*"weight" + 0.335*"loss" + 0.246*"lose" + -0.182*"diet" '
  '+ -0.173*"fat" + -0.109*"exercise" + 0.073*"good" + -0.064*"diabetes" + '
...
  '0.085*"say"'),
 (19,
  '0.612*"people" + 0.474*"time" + 0.306*"know" + 0.218*"say" + -0.215*"need" '
  '+ 0.171*"want" + -0.154*"make" + 0.107*"see" + 0.099*"ass" + 0.097*"think"')]
```

Fig. 2: Topics from topic modeling

By calculating the amount of semantic correspondence between high scoring terms in the topic, Topic Coherence measures the score of a single topic. These metrics assist in separating issues that can be understood semantically from topics that are outcomes of statistical inference.

- Initially no. of topics provided to the model in LDA is 20 which gives a coherence score of 0.3475.
- LDA shows coherence score of 0.5473 with 222 optimal no. of topics.
- Initially no. of topics provided to the model in LSA is 20 which gives a coherence score of 0.3495.
- LSA shows coherence score of 0.4132 with 222 optimal no. of topics as shown in Fig 4.

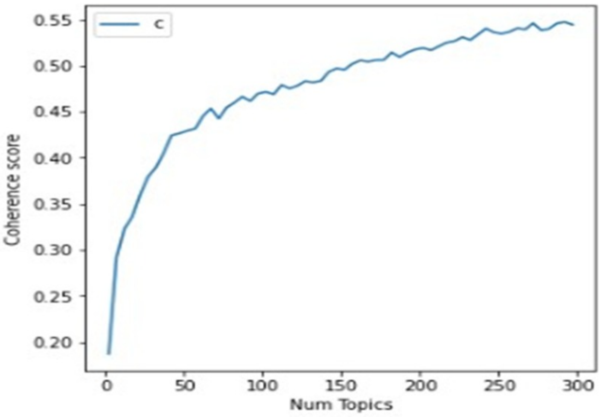


Fig. 3: Coherence score vs Num of topics in LDA

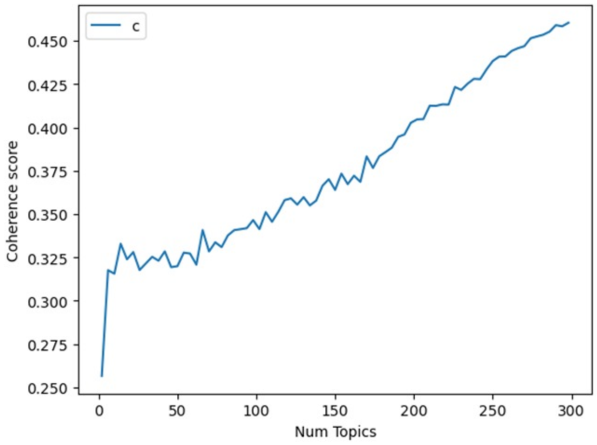


Fig. 4: Coherence score vs Num of topics in LSA

Table 2 DDEO topics and subtopics – diabetes, diet, exercise, and obesity are shown

Topics	Subtopics
Diabetes	Diabetes type2 Obesity Diet Blood Pressure Alzheimer
Exercise	Fitness Obesity Daily plan Diabetes
Diet	Obesity Weight loss Diabetes
Obesity	Diet Alzheimer Exercise Diabetes

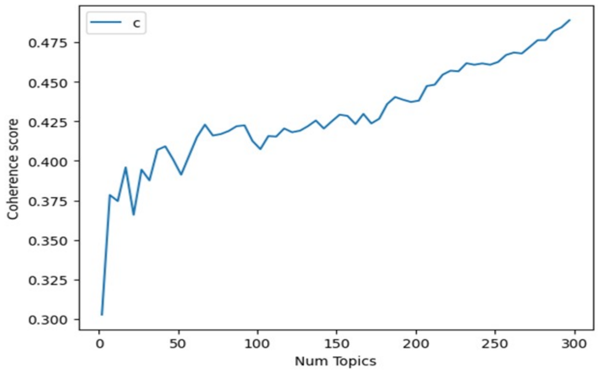


Fig. 5: Coherence score vs Num of topics in LSA TFIDF

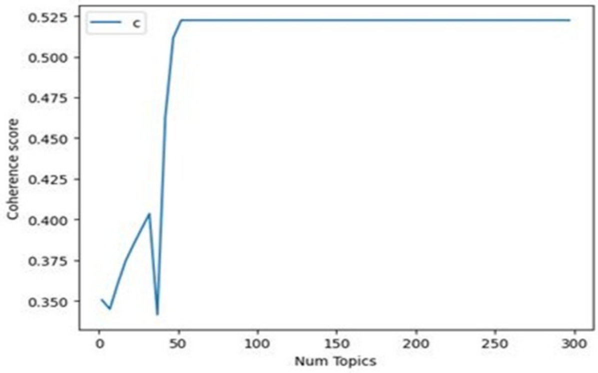


Fig. 6: Coherence score vs Num of topics in LDA TFIDF

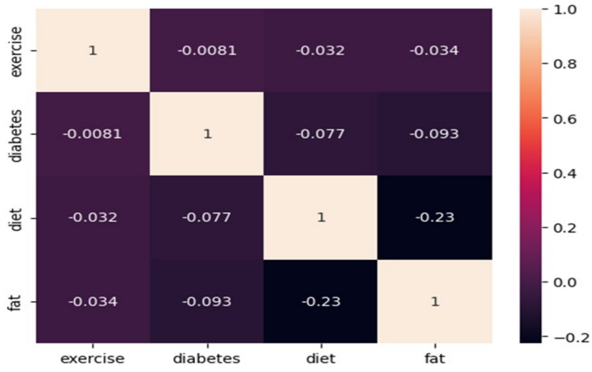
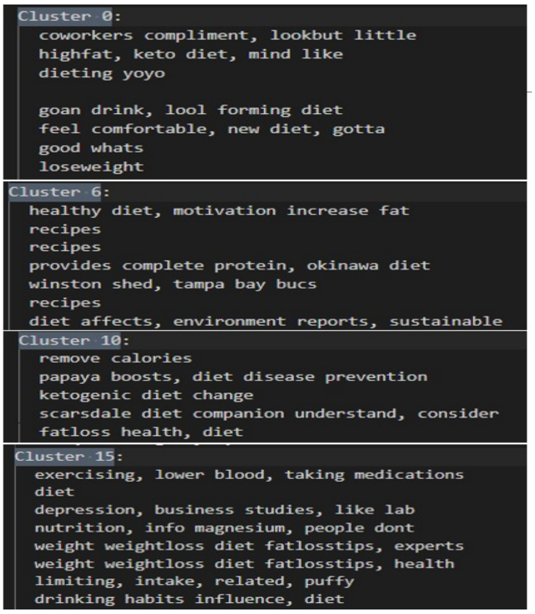


Fig. 7: Correlation plot between the topics



```
Cluster 0:
  coworkers compliment, lookbut little
  highfat, keto diet, mind like
  dieting yoyo

  goan drink, lool forming diet
  feel comfortable, new diet, gotta
  good whats
  loseweight
Cluster 6:
  healthy diet, motivation increase fat
  recipes
  recipes
  provides complete protein, okinawa diet
  winston shed, tampa bay bucs
  recipes
  diet affects, environment reports, sustainable
Cluster 10:
  remove calories
  papaya boosts, diet disease prevention
  ketogenic diet change
  scarsdale diet companion understand, consider
  fatloss health, diet
Cluster 15:
  exercising, lower blood, taking medications
  diet
  depression, business studies, like lab
  nutrition, info magnesium, people dont
  weight weightloss diet fatlosstips, experts
  weight weightloss diet fatlosstips, health
  limiting, intake, related, puffy
  drinking habits influence, diet
```

Fig. 8: Bert embedding topics

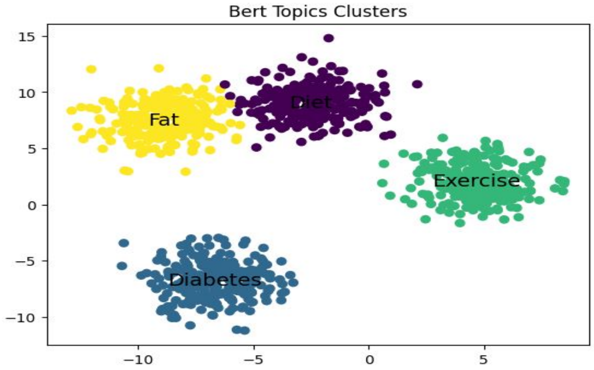


Fig. 9: Bert Topics Clusters

The strongest correlation was found between topics Diet and Obesity (-0.23) and lowest correlation between Diabetes and Exercise(-0.0081).

As the results obtained in Table 3 LDA performs better than LDA-TF-IDF because LDA considers the co-occurrence of words within documents and across documents, whereas LDA-TF-IDF only considers the frequency of words within documents. This means that LDA can identify topics that are more coherent and have a stronger semantic meaning than LDA-TF-IDF. Similarly, LSA-TF-IDF performs better than LSA because LSA-TF-IDF considers both the word's frequency in the document and its inverse frequency in the corpus. This suggests that compared to LSA, LSA-TF-IDF can find subjects that are more coherent and have deeper semantic significance.The topics that are generated by bert models are much more interpretable than other models.As shown in Fig. 9 topics are fat and diet are much more correlated which can be seen in bert embedding topics in Fig. 8.

Table 3 Coherence Score Comparison Table

Modelling Technique	20 topics	222 topics
LDA	0.3475	0.5473
LSA	0.3495	0.4132
LDA TFIDF	0.3847	0.5224
LSA TFIDF	0.3659	0.457

V. CONCLUSIONS

LDA gave a better coherence score for the same number of optimal topics then LDA-TF-IDF, LSA and LSA-TF-IDF because LSA was mostly concentrated on the dimensionality reduction and LDA-TF-IDF only considers frequency of words within documents.Also the bert embedding model generates topics that are much more interpretable and can visualized in better way. The correlation between these topics show the strongest correlation between the topics diet and obesity. The lowest correlation was found between topics diabetes and exercise.

REFERENCES

Arnold, C. and Speier, W. (2012). A topic model of clinical reports. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1031–1032.

Dahal, B., Kumar, S. A., and Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20.

Edo-Osagie, O., De La Iglesia, B., Lake, I., and Edeghere, O. (2020). A scoping review of the use of twitter for public health research. *Computers in biology and medicine*, 122:103770.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.