Topic modelling in precision medicine with its applications in personalized diabetes management

Ni Ki, C., Hosseinian-Far, A., Daneshkhah, A. & Salari, N.

DOI 10.1111/exsy.12774
ISSN 0266-4720
ESSN 1468-0394

Publisher: Wiley

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH PAPER

Topic Modelling in Precision Medicine with its Applications in Personalised Diabetes Management

Ni Ki Chong[1] | Amin Hosseinian-Far[2] | Alireza Daneshkhah*[3] | Nader Salari*[4]

[1]School of Computing, Electronics and Mathematics, Coventry University, Coventry, UK
[2]Centre for Sustainable Business Practices, University of Northampton, Northampton, UK
[3]Research Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK
[4]Department of Biostatistics, Kermanshah University of Medical Sciences, Kermanshah, Iran

Correspondence
*Corresponding authors; Email:
n.salari@kums.ac.ir;
ac5916@coventry.ac.uk

Summary

Advances in Internet of Things (IoT) and analytic-based systems in the past decade have found several applications in medical informatics, and have significantly facilitated healthcare decision making. Patients' data are collected through a variety of means, including IoT sensory systems, and require efficient, and accurate processing.Topic Modelling is an unsupervised machine learning algorithm for Natural Language Processing (NLP) that identifies relationships and associations within textual data. The application of Topic Modelling has been widely used on raw text data, where meaningful clusters (topics) are generated by the model. The purpose of this paper is to explore the varying methods of Topic Modelling, mostly the Latent Dirichlet allocation (LDA) model, and its applicability on personalised diabetes management. The proposed study evaluates the possibility of applying topic modelling methods on diabetes literature and genomic data in order to achieve precision medicine.

KEYWORDS:
IoT-based systems for healthcare, Topic modelling, Latent Dirichlet allocation, Personalised diabetes management, Precision medicine, NLP

1 | INTRODUCTION

Whilst the practice of personalised medicine is not new, data-driven healthcare can be considered as the future of medicine. Precision medicine is an initiative that considers the varying genetic make-up of individuals when deciding on a treatment and/or prevention method. The study of the human genome and its environmental factors allow pharmaceutical experts to predict an individual's response to a certain drug and/or treatment and thus improving the quality of care received by the patients. Primarily, this research aims to dive deeper and harvest the potential of novel machine learning, known as topic modelling [1], in the of understanding a complex disease (diabetes), by analysing the related text data available in different formats and lengths. Diabetes is notoriously complex, which makes its prevention and treatment a challenging task. Due to the successes of topic modelling that have been observed in other research areas [2, 3, 4, 5], there is an untapped potential in the use of this method when dealing with diabetes text data collected from various sources. There is a clear room for improvement in the diabetes management process, thus this paper addresses the characteristics of diabetes type 1 and 2, and the current advances in the personalised management of diabetes treatment by utilising the topic modelling approach. Based on the collected data sources from the IoT-based systems and developed data-analytic methods in this study, the proposed tool would have the potential in achieving a better understanding of the diabetes type 1 and 2 through the invaluable information intelligently extracted from the text data, providing diabetes researchers with the knowledge that could hopefully be built upon for improved diabetes management.

The paper is structured as follows: Section 1 provides a thorough and detailed review of a variety of topics in this multidisciplinary study. Section 2 outlines the Topic model that was implemented within this study for diabetes management. Section 3 presents the results, followed by a critical discussion in Section 4. The paper is bookended with the summary of key points within the conclusion section.

1.1 | A Background on Health Informatics, and Precision Medicine

1.1.1 | Drug Discovery

Drug discovery has always been a lengthy process to go through and has been averaged to take around 10-15 years to get a drug onto the market [6]. This process is not only time consuming, but is at the same time incredibly inefficient. Jang et al. [7] address this problem with topic modelling. Jang et al. proposed an algorithm that models journal abstracts on gene-drug relationships. This was then mapped with a drug-side effect relationship in order to identify the plausible side effects associated with a particular gene and drug. The proposed algorithm learns from each document it is fed, continuously learning from every interaction that it has with the data. As a classifier, PISTON was able to out-perform other co-occurrence classification models considerably. With this algorithm, a classifier for genetic characteristics was built – which could accurately identify drugs that would have a positive outcome or no side effect (at the bare minimum) on the gene candidate. This study demonstrates the successful usage of Topic Modelling whereby textual data such as journal extracts exhibit potential in being a credible source for hidden knowledge on for medical purposes.

1.1.2 | Identifying associations of genetic variants and diseases

Over the last decade, several studies [8, 9, 10, 11] explored the usefulness of data analytic methods and visualisation tools to model and simulate the development of malignant tumors. They attempted to illustrate how these tools and methods can be helpful in identifying weak spots of tumor for treatment, inspecting malignant tumors in general and inspecting whether genes have cancerous cells.

Another effective tool for the personalised healthcare is the recommender system [12] which has been very useful in analysing the biological data for gene prediction, and recommending the genes for individual patients based on the identified the gene interest (Gi). This can be done using the TOP-N Gene-based Collaborative Filtering (GeneCF) algorithm developed around the Gi of patients. The GeneCF algorithm can assist doctors to provide smarter, and more customized care for cancer patients.

A study by [13] highlights the usage of electronic health records (EHRs) in modelling the associations between genetic variants and disease phenotypes. The study was mainly interested in modelling the increased risk of cardiovascular diseases and hyperlipidaemia that is typically associated with a single nucleotide polymorphism (rs10455872 in LPA). EHRs were modelled using non-negative matrix factorization (NMF), i.e. another type form of topic modelling, to extract relevant topics from the phenotype data. Among the 12759 individuals studied, cardiovascular disease was prevalent in 33% of the entire cohort. The topics were found to be consistent with the theory of association of disease between certain phenotypes. Results were also compared against Phenome Wide Association Studies and clinical experts were consulted to ensure that the results were conclusive. Interestingly, the topic model was able to discover an association between the rs10455872 LPA variant and lung cancer, an unidentified association prior to this study.

1.1.3 | IoT-based systems for personalised diabetes management

Over the past decade, advances in Internet of Things (IoT) technologies and devices, have resulted in streamlining precision medicine data capture and personalised health systems [14, 15]. Reduction in price of IoT devices, improvement of their computational efficiency, and their enhanced accuracy of operation have played a significant role in these new developments [16]. [17] proposed an IoT-based smart hospital system which relied on a combination of radio frequency identification (RFID), wireless sensory networks (WSN), and mobile system. Later, [18] argued that one of the key barriers for the expansion of IoT-based systems for healthcare is their interoperability, and the heterogeneous nature of systems and software. Other researchers have also highlighted security challenges and potential issues with respect to personalised IoT-based health devices [19]. The application for technology was not primarily based on the overall hospital and healthcare units, but found its way in personal healthcare devices, interacting with patients; this is even more beneficial for managing chronic diseases [20]. [21], in a recent study, reviewed several existing research works and applications of IoT and wireless personalised diabetes management systems. Scholars have proposed several innovative and interesting applications of IoT-based systems for managing diabetes. Instances of these include, and IoT based device for the diagnoses of diabetic ketoacidosis by testing breath of patients [22], a system for Urine-based Diabetes by mentoring patient's urine [23], mobile IoT system for non-invasive glucose level sensing [24], use of sensors to measure glucose level in the blood, among others.

It is also of great importance is to ensure the security of patients' healthcare data, to implement access control for normal and emergency scenarios, and to support smart deduplication to save space in the big data storage system. Recently, several privacy-preserving smart

IoT-based healthcare big data storage systems with self-adaptive access control have been proposed (e.g.,[25]). Using these systems, medical files generated by the healthcare IoT network are encrypted and transferred to the storage system, which can be securely shared among the healthcare staff from different medical domains leveraging a cross-domain access control policy. In addition to the privacy and security of the smart healthcare systems, the ethical issues of using the smart IoT-based healthcare and systems and devices should be also considered. [26] suggested that vulnerabilities of the system which may pose potential threat to users should be first identified. The misused data which can also cause negative impact must be prevented, and companies should be more transparent on the usage of personal healthcare data. It is thus crucial to improve the users' awareness and knowledge to protect their personal data.

1.1.4 | Clinical Decision Support

In a study by [27], automating the ordering of clinical order sets based on hospital records and past order sets was proposed. The model was targeted clinicians and was to provide them with the opportunity to make evidence-based decisions. Traditionally, clinicians follow the general process of searching up a disease on the database and would have to make decisions on the treatment for the patients based on past experiences and the electronic health records. They proposed an algorithm to identify a structure within these clinical order data in predicting decisions. Their approach utilises the LDA topic model in identifying the topics. Topic Modelling was concluded to be able to identify thematic structures within clinical data such as order set data. The potential use case for this is minimizing the human-input and room for human-prone error. Naturally, algorithms will be unable to completely replace the role and expertise of a clinician as the role requires expert decision-making. However, topic models can act as a support system for a clinician, to provide basis for their decisions that will ultimately be up to their own expertise.

1.1.5 | Adverse Drug Reactions

The burden of adverse drug reactions can affect a patient's well-being as well as the healthcare provider. When a drug is being tested for the market, this typically involves a relatively small- scale testing trial, compared to the immense scale of patients that will actually be using the drug. [28] recognised the problem and using topic modelling methods proposed to offer a data-driven method to provide basis on which a drug should or should not be consumed with the risk of a negative reaction. The solution applies LDA to uncover hidden topics within the drug structure that is causing specific adverse reactions. Topic modelling – often times found in a commercial context has been picking up traction within the medical domain. Its usage has proven to be versatile in this section, from discovering new uses of existing drugs, to providing support to medical practitioners in decision making. It is evident that Topic Modelling is prevalent within the domain and has the potential to be further developed for the purposes of diabetes journals.

1.1.6 | Using Electronic Medical Records

The typical patient diagnosed with T2D would be advised to maintain a healthy lifestyle through healthy eating and regular exercise. This is accompanied by their prescribed medication and/or insulin injections [29]. While these general guidelines can only be effective on surface-level and/or less severe cases. In 2012, the American Diabetes Association recognised the need for a more personalised management system, in order to ensure that patients receive the ideal care for them. Phenotypes and genetics play a huge role in the responses to treatment, which is demonstrated in a study done by [30]. It was demonstrated that African Americans showed a better response towards metformin as opposed to Caucasian Americans with similar prediabetes characteristics. As there are evident subgroups within the population of affected individuals, it is essential that we address the biological differences when it comes to receiving medical treatment. [29] also introduce an algorithm that incorporates the use of electronic medical records and data analysis to provide patients with a case by case level of care.

Through the usage of actual medical records provided by the Boston Medical Centre, this study was driven by actual patient data, observed over a year. Patient medical histories included hospital visits; the types of therapies prescribed based on the current situation of the patient at the point in time. The demographic data of each patient allowed for categorization of the type of individuals. The combination of their medical history as well as demographic data provided insight on the suggested therapies for future patients. Each patient's visit and therapy history were used to train the k-nearest neighbours' algorithm, in order to predict the future patient's next line of therapy. Based on the independent variables, the therapy with the best predicted outcome was suggested to the future patient. This study concludes the benefits that arise from personalised diabetes management, yielding an improved outcome in HbA1c test compared to the unit standard.

1.1.7 | Improved Glycaemic control

Through the standard care and treatment for diabetes, many patients are underachieving the goals for their planned treatments. The management of diabetes can be complex, and often times people find themselves struggling with having to self-adjust to their treatment dosages in their everyday life. The design of Integrated Personalised Diabetes Management [31] involves a process of further testing in order to understand the patient's current HbA1c levels. Following this with every visit, doctors would be able to recommend a structured therapy and monitor the progress of each patient with data collection and analysis. This would allow the doctor to gain structured insight and monitor the patients progress levels throughout therapy. Creating personalised profiles of each patient allowed medical practitioners to make decisions based on the individual characteristics as opposed to the traditional one-size- fits-all approach. A ripple effect seen during this study was that patients seem to have built a trusting relationship with their doctors, improving the overall treatment for the patient. The outcome of this study showed an overall improvement in the control of their glucose levels. With the vastly growing database of medical health records, the treatment of diabetes is heading towards the direction of individualization in order to amplify the efforts in treating the disease.

1.2 | Natural Language Processing in Medicine

1.2.1 | Knowledge database

Statistical methods for NLP are widely used due to the broad nature of language. A language can contain multitudes of rules and formats that statistically modelling a passage became far more efficient over the traditional linguistics way of encoding the 'laws of language'. The differentiation between linguistics-based approaches and statistical based approaches is in how the machine views the document. Through statistical methods, a document is typically represented in a matrix format. On the other hand, linguistic methods are where experts embed the rules of language by enforcing rules and knowledge within the domain [32]. [33] proposed a system – GENIA where it simplifies the task of text annotation in the biomedical domain. The system incorporates both statistical and linguistic methods when modelling language, by defining linguistic rules as well as knowledge within the domain. The task focuses on the molecular information within journal abstracts, which successfully overcame the limitation of pure statistical NLP methods. Especially within an industry as critical as medicine, it is crucial for algorithms and methods to incorporate domain-specific knowledge into the solutions, in order for essential texts or words not to be lost in translation.

In 2019, Tran et al. [34] reported that although AI–based methods have been rapidly developed in the last two decades with the vast applications in medicine and health care, there is a lack of comprehensive reporting on the productivity, workflow, topics, and research landscape of AI in this field. As a result, they have used the LDA-based NLP approach, which is very similar to the method used in this paper. The methods was applied to evaluate the global development of scientific publications and constructed interdisciplinary research topics on the theory and practice of AI in medicine from 1977 to 2018. They have concluded that applications of AI have mainly impacted clinical settings, data science and precision medicine (collecting individual data for precision medicine), and policy making. However, the precision medicine applications have not been commonly used in resource-poor settings due to the limit in infrastructure and human resources. Therefore, it is necessary to conduct further research to examine and incorporate evaluations of implementing AI-based solutions, precision medicine and their contribution to health system sustainability [35].

1.2.2 | Structured and Unstructured data

When dealing with electronic health records, often times records or entries will contain words that can exist in several variations such as cardiac arrest, heart attack, cardiovascular disease and many more. These variations make modelling and representing data for analysis complex and less straightforward. [36] highlights the issue by testing a hypothesis on a collection of 2,000 cardiology notes. These notes were collected from a collection of notes that were initially dictated by a cardiologist, and then translated into electronic text automatically. It was noted that the dictated texts resulted in the loss of some words during the process of translation, and the question on what the value of the missing words arose. The study brings to light the need for structured data and the standardization of certain terms for optimal processing. Similarly, in medical journals, there is the tendency to abbreviate longer or relatively common terms to shorten the passage. Understandably, there is a need for shorter and abstract terms in order to maintain readability. However, this leads to other problems when using statistical NLP to model textual data (see also [37] and references therein for further relevant discussion).

1.2.3 | Resource limitation

One of the critical challenges when dealing with data in medicine is privacy. Data collection is also a key challenge, and typically the biggest one in the context. The nature of big data is complex, commonly unstructured and requires processing in order to extract knowledge from it. Adoption of big data methods or systems come with a major problem of privacy. Unlike other sources of data, medical data come with a high level of sensitivity. Understandably so, the restrictions placed on access to medical data is incredibly high. A huge driver of machine learning algorithms is the availability of data. Without freely available data, it inhibits the capacity for researchers to perform analysis in order to improve the efficiency of current and future healthcare systems. Nevertheless, there are organizations dedicated to the controlled resource sharing of medical data. Novartis is a medical data sharing association that transparently allows patients to share their medical data for clinical trials and studies (see [38] for further details about this association). Similarly, with Topic Modelling methods, it is critical for the model to have relevant and reliable data. In doing so, models can be improved, and systems can be implemented without the possibility of computational mistakes. With learning algorithms, accuracy improves as the amount of data increases. That said, it is important that data governance measures are in place in order to protect the rights and privacy of individuals.

There have been efforts in penetrating the healthcare sector with NLP methods, with databases dedicated to texts in the medical domain. As we move towards automation, the amount of textual data available in healthcare is immense. Case study data, medical reports, medical histories, are all examples of possible data that could be gathered and modelled with NLP. The hidden structures and information in texts have been demonstrated within the various other industries commercially and it can be assumed that there is untapped potential within healthcare context as well.

1.3 | Research gaps

While the concept of topic models is not new, limited studies were conducted on topic models for finding the hidden structures within diabetes textual data. Moreover, the extent of which topic models that have been applied within the healthcare sector was limited to identifying adverse drug reactions, clinical decision support, and genetic associations. Diabetes has been notorious for being complex and it would be certain that the treatment for it would follow in the same manner. Identifying possible phenotype and genotype associations can be beneficial in the treatment of the disease. Hence, this study aims to answer the question of using topic modelling for the purpose of uncovering hidden structures within textual data. The intended data used for this application is case study data, whereby in-depth studies are done on a population. For instance, the study of a certain diabetes treatment and the reaction a population sample would have towards it. This could take form as a particular gene that a population shares, and their reaction to a certain drug. While the data available might not be enough to deduce a specific result, PubTator contains resources that emulate a similar idea, without having to obtain sensitive information from individuals.

1.4 | Understanding the Disease

Diabetes mellitus is a complex disease known to hinder the secretion of insulin in the body. This phenomenon has been widely attributed to environmental as well as genetic factors. According to the World Health Organization, there are over 422 million patients diagnosed with diabetes. This amounts to 1 in every 11 people being affected by this disease.

1.4.1 | Type 2 Diabetes (T2D)

T2D is the most prevalent type within the population – affecting over 91.2% of diagnosed diabetes patients in the United States alone [39, 40]. This form of diabetes is when the body is resistant to the insulin produced or is unable to produce enough for the body. People affected by this disease would be also susceptible to a list of recurring health problems in the future such as cardiovascular diseases, nerve damage, kidney damage, and many more [41]. There have been strong evidences suggesting that 20%-80% of diabetes susceptibility is associated with the population, and family genetics [42, 43]. It is known that genetics plays a role in the susceptibility of diabetes, however, there has been no direct linkage with a particular gene and the disease. That said, there is a list of genes that do contribute to the disease susceptibility.

Table 1 provides an example of the type of genes associated with T2D, and how it affects the body. It is worth noting that gene-environment interactions also play an important role in the susceptibility of the disease. Lifestyle factors can influence the genes in insulin production, as well as the breakdown of glucose. With the extensive list of genetic associations and its mutations combined with the environmental associations, the efforts in treating the disease become more extensive in parallel.

TABLE 1 Genes contributing to Diabetes [44].

Gene	Description
Hepatocyte nuclear factor-1A (HNF1A)	This gene and its other mutations are the genes commonly found in young diabetic patients. It affects the development of liver and proper functioning beta cells.
Calpain-10 (CAPN10)	Found to modify the receptor signaling of intercellular functions of the body. Known to be truly associated with T2D.
Transcription factor 7-like 2 (TCF7L2)	Seen as a high-risk gene for diabetes susceptibility. The association has been identified to the single-nucleotide polymorphisms (SNPs) within the genetic locus itself.

1.4.2 | Type 1 Diabetes (T1D)

Type 1 diabetes is whereby the pancreas produces little to no insulin for the body (World Health Organization (WHO)-Diabetes). This rare type of diabetes affects a smaller percentage of the population but is more genetically dependent. The disease is caused by the destruction of cells in the pancreas, inhibiting the production of insulin [45]. While the genetic component is prevalent in T1D, for the purpose of this study, we will be focusing on modelling T2D datasets due to the larger volume of available case study data.

1.4.3 | Advances in Diabetes Management

With diabetes management, the decision process for treatment can be broken down into many factors. According to an article from the American Diabetes Association, in 2014, the field began with what was called “Individualised Therapy” for diabetic patients [46]. This form of therapy strays from the standard forms of treatment where patients are treated as one-size-fits-all when receiving treatment. Since then, there has been strides in the community where machine learning methods have slowly made its way into the conversation. Previously, practitioners took into consideration the patients age, their glucose control history, even personal characteristics when deciding treatment plans for type 2 diabetes. Now with the extensive EMR availability, the possibilities for improved personalised treatments is tenfold.

Recent studies suggest the prospect of applying machine learning methods to EHRs in personalised medicine. A paper on the personalisation of diabetes care highlights the usage of the classification technique – k-nearest neighbours [29]. The algorithm maps training examples in a feature space and classifies new input based on its nearest feature neighbour [47]. The output of the proposed model was a treatment recommendation based on certain features of a patient (sex, race, BMI, treatment history, age, and diabetes progression). The paper proposes a prescriptive algorithm that generates a recommendation dashboard for the patient based on their current regimen. In a similar study conducted by Vu et al. [48], they have presented an inclusive landscape of application of AI in diabetes through a bibliographic analysis and offers future direction for research. They combined bibliometrics analysis with exploratory factor analysis, and utilised the LDA model to explorer emergent research domains and topics related to AI and diabetes.

1.4.4 | Precision Medicine

Studies have shown that in certain diseases, genetic variability contributes to a large part of how an individual may respond to a particular drug or treatment [49]. Furthermore, it can also determine the patient’s susceptibility to a particular disease such as diabetes. Though a recent initiative, precision medicine has been in practice for several years now, a common practice is blood transfusions.

In 2015, the then-president Barack Obama announced a brand-new healthcare initiative - precision medicine. The initiative aims to tackle diseases and preventive care by characterising patients by their genetic variability. With its initial approach to target oncology, this initiative utilises the ever-growing database of human genomic sequences, coupled with increased computational intelligence creates a well-informed basis for clinical decisions. Since then, precision medicine has been implemented on numerous accounts, such as Intermountain Healthcare [50]. Intermountain Healthcare is an integrated healthcare delivery system that offers genomic testing to patients under their program. This prior analysis allows clinicians to understand the patient’s needs better, as well as provide patients with an informed treatment plan. The results of this program have provided patients with better clinical outcomes at a similar, if not lower cost. Due to the successes of this program, Intermountain has modified its own policies in favour of this new initiative, placing more confidence in it as success grows. Similarly, the Levine Cancer Institute [51], houses an electronic clinical pathway system that services oncologists in its community. This system uses the patient's genetic reports to find a suitable cancer treatment trial, and if none, it will find the suitable treatment for them that exists based on their genetic reports. The system connects 26 cancer sites across the United States, which means it has access to a large genomic database for clinicians to rely on.

1.5 | Topic Models

The intuition behind topic models is to tackle the difficult task of summation, especially when it comes to large volumes of text. Fundamentally, topic models aim to statistically determine the topic of a document by evaluating word/topic probabilities within a document. Since the introduction of Topic Models, there has been numerous variations of topic models, all ultimately achieving the same goal. In this section, we introduce the most popular model - LDA, Latent Semantic Analysis (LSA), and Probabilistic Latent Semantic Analysis (PLSA), to provide a basis of intuition on how popular topic models work.

1.5.1 | Latent Dirichlet Allocation

Introduced by [1], LDA is a probabilistic topic model that represents a corpus (collection of documents). Generally, topic models tend to assume that a single document (instance) will only fall under a single topic categorization. However, it was found that in most real-world instances, documents tend to encapsulate more than just a single topic. LDA recognises this and assumes that a document is generated with a combination of topics within a single document. By identifying the generative process of a document, the probabilities of topics contained in them can be deduced. It is worthwhile illustrating the intuition behind LDA in order to provide a more intuitive explanation. Since LDA is an unsupervised learning algorithm, it automatically detects the topic distribution within a corpus based on the defined parameter (number of topics). This parameter will be defined manually and will be the basis for which the corpus is modelled. In order to grasp the intuition behind LDA, we have to dive deeper into the document generation process.

With the defined number of topics (in a corpus), the topic distribution in each document will be sampled from the defined number of topics. For example, a corpus has three topics in total: COVID-19 (see e.g., [52, 53, 54, 55] for some recent systematic reviews and predicting mortality rate of this infectious disease), Diabetes, and Cancer. Within each topic is a word distribution that describes a particular topic. Similarly, with the human cognitive phenomena, when reading consuming text, humans tend to pick up key words that are associated with a specific topic. This is where the word distribution comes into effect. When deciding what topic, a sentence or paragraph belongs to, the word distribution on the topic of diabetes that would likely to have the words 'glucose' or 'insulin' in them. Certainly, this example has been severely understated and the number of words within the word distribution is far larger than just 'glucose' and 'insulin', yet, it exemplifies the idea behind the topic-word distribution.

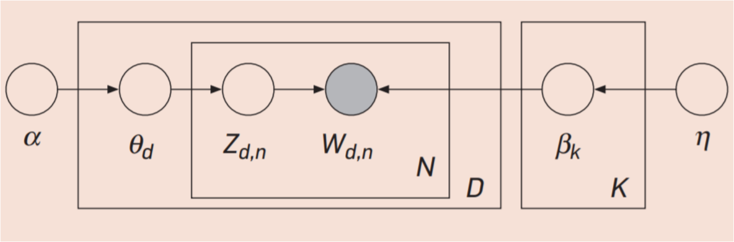


FIGURE 1 Graphical representation of LDA document generation (adapted from [1]).

The LDA model, as illustrated in Figure 1, can be formally described with the following notation. The topics are denoted by $\{\beta_k : k = 1, \ldots, K\}$, where each β_k is a distribution over the vocabulary (or the distributions over word). The topic proportions for the dth document are $\theta_d = \{\theta_{d,k} : d = 1, \ldots, D;\ k = 1, \ldots, K\}$, where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for the dth document are $Z_d = \{Z_{d,n} : d = 1, \ldots, D; n = 1, \ldots, N\}$, where $Z_{d,n}$ is the topic assignment for the nth word in document d. Eventually, the observed words for document d are denoted by $W_d = \{W_{d,n} : d = 1, \ldots, D; n = 1, \ldots, N\}$ where $W_{d,n}$ represents word assigned to nth word in document d, which is an element from the fixed vocabulary. (see [1, 56] for further details about topic modelling).

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \tag{1}$$

It should be noted that the distribution given in Eq. (1) determines a number of dependencies. For instance, the topic assignment $\mathsf{z_{d,n}}$ depends on the per-document topic proportions θ_d, defined above. Furthermore, the observed word $\mathsf{w_{d,n}}$ depends on the topic assignment $\mathsf{z_{d,n}}$ and all of the topics $\beta_{\mathsf{1:K}}$. From the practical perspective, that term is defined by looking up as to which topic $\mathsf{z_{d,n}}$ refers to and looking up the probability of the word $\mathsf{w_{d,n}}$ within that topic (see [1] for further details).

We now briefly discuss the generative process for LDA here, as illustrated in Figure 1 and given in Eq. (1). The assumptions prior to the generation process is that both the Dirichlet priors η and α are determined. The process begins with the assumption that there are K topics within the corpus. The topics K is then distributed based on the Dirichlet prior α, across the set of documents D. Each word is then generated based on the process outlined in Figure 1. This eventually generates a set of words, N. We then assign a topic $\mathsf{Z_{d,n}}$ for each of the generated words in set N. Using the Dirichlet prior η, and the topic assumption $\mathsf{Z_{d,n}}$, a word is generated $\mathsf{W_{d,n}}$ This iterative process is repeated across the entire corpus till it converges [1].

The assumption that documents generated are based on a pre-determined set of topics suggests a high suitability for the task at hand. The nature of each document extracted in this context aligns with the base assumption of LDA. For example, an abstract within the dataset used for this investigation covers the gene relationship and how it affects diabetic patients. Within the same document, it also covers areas of cancer, and its correlation with the gene, as well as diabetes.

Here, we discuss a simple example that how the topic distribution in each document will be sampled from the defined number of topics when a LDA-based topic modelling was fitted to the defined number of topics (in a corpus). Let assume, there are three topics in total in the corpus: COVID-19, Diabetes, and Cancer, as illustrated in Figure 2. Within each topic is a word distribution that describes a particular topic. Similarly, with the human cognitive phenomena, when reading consuming text, humans tend to pick up key words that are associated with a specific topic. This is where the word distribution comes into effect. When deciding what topic, a sentence or paragraph belongs to, the word distribution on the topic of diabetes would likely to have the words 'glucose' or 'insulin' in them. Indubitably, this example has been severely understated and the number of words within the word distribution is far larger than just 'glucose' and 'insulin', but it exemplifies the idea behind the topic-word distribution.

1.5.2 | Latent Semantic Analysis

Similarly, with LDA, Latent Semantic Analysis is a probabilistic topic model. Document instances are represented as a matrix format whereby rows are the unique words and columns are the occurrence in the corpus. The cells of the matrix indicate the word frequency within the documents. LSA characterises topics based on the semantic space of the document, assigning higher similarity for words that appear closer together.

In order to reduce the dimensionality of the matrix, singular value decomposition (SVD) is applied on the matrix, that identifies the unique cells (words) that are in the matrix [57]. In essence, LSA is the product of the matrices of word-topic probabilities, the topic importance, and the topic distribution across the documents.

$$A \approx U_t S_t V_t^T$$

Intuitively, LSA assumes that the weight of a word higher when it frequently occurs in a single document, and infrequently in the entire corpus. This is where Term Frequency-Inverse Document Frequency method comes in to form a word-document matrix

Based on Figure 3, U represents the word assignments to topics of terms (rows) and topics (columns). Vector V represents topic distribution over the documents, of topics (rows) and documents (columns). Vector S is the importance of topics across documents, of topic importance (rows and columns). Vector S is a diagonal matrix with cells with $\mathsf{0}$s except for the diagonal cells containing the topic importance.

1.5.3 | Probabilistic Latent Semantic Analysis (PLSA)

This popular form of LSA has gained traction over the years for its probabilistic nature in modelling documents. Unlike LSA, PLSA treats documents as a combination of multinomial distributions [59]. In the same way as LSA, PLSA uses SVD in order to break down the dimensions of the matrix. The primary difference between PLSA and LSA is the cells within the matrices. In PLSA, the matrices contain the document, word, and topic probabilities in replacement of the document, word, and topic assignments found in the LSA method.

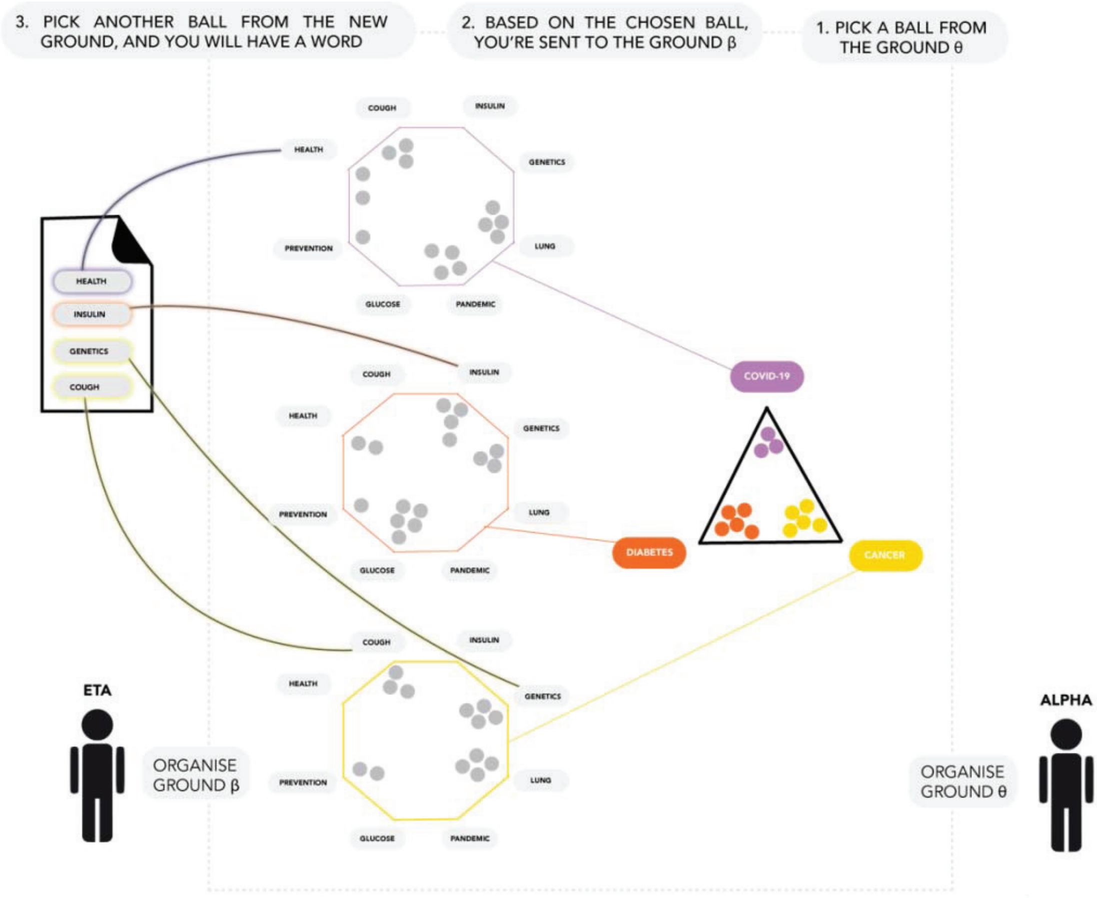

3. PICK ANOTHER BALL FROM THE NEW GROUND, AND YOU WILL HAVE A WORD
2. BASED ON THE CHOSEN BALL, YOU'RE SENT TO THE GROUND β
1. PICK A BALL FROM THE GROUND θ
COUGH
INSULIN
HEALTH
GENETICS
PREVENTION
LUNG
GLUCOSE
PANDEMIC
COVID-19
DIABETES
CANCER
ETA
ORGANISE GROUND β
ALPHA
ORGANISE GROUND θ

FIGURE 2 Visual representation of LDA to select the most relevant topics to Cancer, Diabetes and COVID-19.

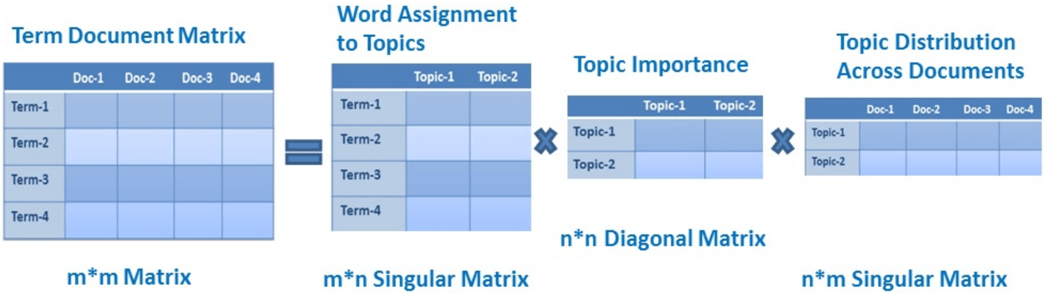


FIGURE 3 Matrix representation of LSA (adapted from [58]).

1.5.4 | Topic Coherence

A challenge when dealing with unsupervised topic models is defining the number of topics for a given corpus. Defining the optimal number of topics when topic modelling is a crucial step in achieving an appropriate word-topic distribution of the corpus. That said, researchers recognise this issue and have proposed various methods in identifying the optimal number. In 2009, a measure of topic coherence was introduced whereby a word not belonging to the topic will be included in the set of words for a respective topic, and participants (human users) were tasked to identify the 'intruder words' [60]. The intuition behind this is the assumption of the higher the topic coherence, the more likely it is to identify words that did not belong to the word set. Research has shown that this measure outperforms any other measure in topic evaluation, however, human participation can be costly. Aside from the human-cost of evaluating topics this way, human-based coherence measures become a lot more ineffective when participants are not domain experts on the datasets and therefore would not be able to identify words that do not belong to the dataset. For instance, when evaluating topic coherence for a dataset of medical documents, a layman would not be able to identify the intruder words because the medical terms would likely to be unfamiliar in the first place.

Automatic topic coherence measures have been widely proposed by a number of researchers, C_{UMass} [61], C_p [62], C_{uci} [63, 64], C_{npmi}, and CA [65]. Among these is a measure considerably more popular than the rest, which is the C_v measure. This particular measure uses a sliding window method, using a normalised C_{npmi} in order to provide a confirmation measure of the top words in a topic [66]. The usage of this coherence score will be further developed in the methodology section later in this paper.

1.5.5 | Model Selection

Over the years, researchers have developed many variations of topic models, however, the foundations of most are LDA and LSA. In view of that, the topic model comparisons were limited to these two. Between the two, studies have shown that LDA demonstrates a better performance rate over the other probabilistic topic models, due to a higher emphasis on dimensionality reduction compared to other topic models [67]. It has also been identified that LDA has shown a relatively higher performance when it encountered new documents, due to its low dimensionality for semantic representation. Additionally, its smaller semantic representation has led to a lower computational demand [68]. For a study of this scale, it is necessary to maintain the computational demands at a lower rate (without hindering performance), due to the limitation of available resources.

2 | MODEL IMPLEMENTATION

2.1 | Data Collection

For the purpose of this study, PubMed abstracts were downloaded from PubTator, PubMed's API for downloading PubMed data. The search strategy for selecting abstracts was to use the keywords and the operator as "gene" AND "diabetes", in order to find journals relevant to the particular gene-disease association. Since the results did not entirely fit in the gene-disease association criteria, the abstracts were individually selected and downloaded. PubTator limits the download capacity to 100 abstracts per download, hence, this process was repeated 15 times in order to obtain a dataset of 1426 documents. Initially, the idea was to model datasets separately based on the genes identified in section 1.4.1. However, this resulted in a relatively small dataset, producing low topic coherence, as demonstrated in Figure 4.

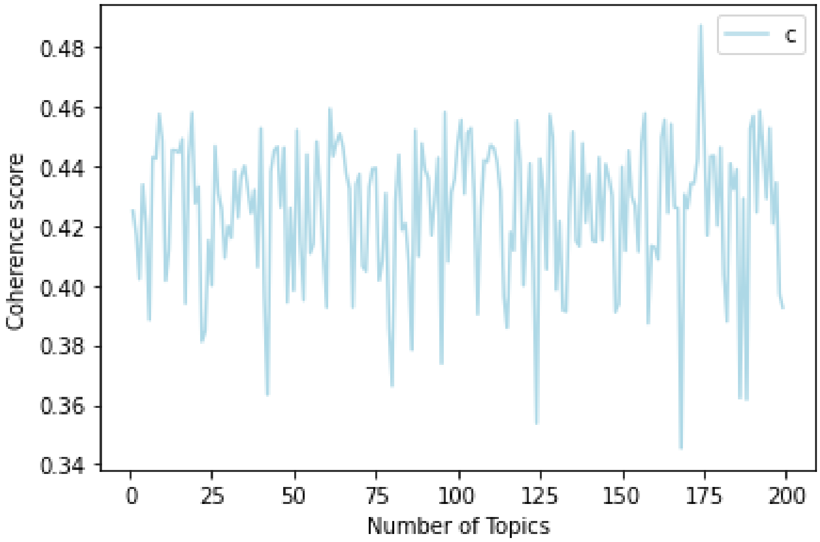


FIGURE 4 Topic Coherence plot for separation of gene data.

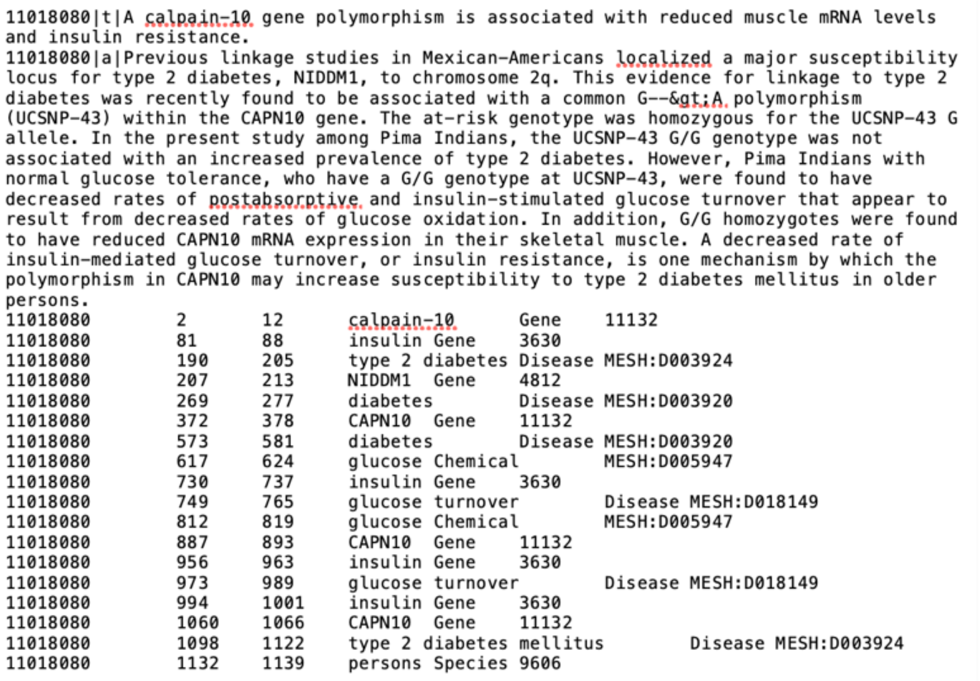
2.2 | Data Preparation

Figure 5 illustrates an example of an abstract downloaded from PubTator. The output typically follows a standard format of:

PMID | t | Title of journal

PMID | a | Abstract

[list of identifiers and tags]



```
11018080|t|A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels
and insulin resistance.
11018080|a|Previous linkage studies in Mexican-Americans localized a major susceptibility
locus for type 2 diabetes, NIDDM1, to chromosome 2q. This evidence for linkage to type 2
diabetes was recently found to be associated with a common G--&gt;A polymorphism
(UCSNP-43) within the CAPN10 gene. The at-risk genotype was homozygous for the UCSNP-43 G
allele. In the present study among Pima Indians, the UCSNP-43 G/G genotype was not
associated with an increased prevalence of type 2 diabetes. However, Pima Indians with
normal glucose tolerance, who have a G/G genotype at UCSNP-43, were found to have
decreased rates of postabsorptive and insulin-stimulated glucose turnover that appear to
result from decreased rates of glucose oxidation. In addition, G/G homozygotes were found
to have reduced CAPN10 mRNA expression in their skeletal muscle. A decreased rate of
insulin-mediated glucose turnover, or insulin resistance, is one mechanism by which the
polymorphism in CAPN10 may increase susceptibility to type 2 diabetes mellitus in older
persons.
11018080        2       12      calpain-10      Gene    11132
11018080        81      88      insulin Gene    3630
11018080        190     205     type 2 diabetes Disease MESH:D003924
11018080        207     213     NIDDM1  Gene    4812
11018080        269     277     diabetes        Disease MESH:D003920
11018080        372     378     CAPN10  Gene    11132
11018080        573     581     diabetes        Disease MESH:D003920
11018080        617     624     glucose Chemical        MESH:D005947
11018080        730     737     insulin Gene    3630
11018080        749     765     glucose turnover        Disease MESH:D018149
11018080        812     819     glucose Chemical        MESH:D005947
11018080        887     893     CAPN10  Gene    11132
11018080        956     963     insulin Gene    3630
11018080        973     989     glucose turnover        Disease MESH:D018149
11018080        994     1001    insulin Gene    3630
11018080        1060    1066    CAPN10  Gene    11132
11018080        1098    1122    type 2 diabetes mellitus        Disease MESH:D003924
11018080        1132    1139    persons Species 9606
```

FIGURE 5 An example of an abstract downloaded from PubTator.

followed by a list of tags and identifiers to categorise the journals. Since we are evaluating the word-document distribution, we do not need to keep the PMIDs and the other tags and identifiers. The first step in data cleaning was to remove the unwanted components in the document. This step was achieved through list manipulation methods in python. Eventually, we end up with a document that looks something similar to:

> “A calpain-10 gene polymorphism is associated with reduced muscle mRNA levels and insulin resistance. Previous linkage studies in Mexican-Americans localised a major susceptibility locus for type 2 diabetes, NIDDM1, to chromosome 2q. This evidence for linkage to type 2 diabetes was recently found to be associated with a common G→A polymorphism (UCSNP-43) within the CAPN10 gene. The at-risk genotype was homozygous for the UCSNP-43 G allele. In the present study among Pima Indians, the UCSNP-43 G/G genotype was not associated with an increased prevalence of type 2 diabetes. However, Pima Indians with normal glucose tolerance, who have a G/G genotype at UCSNP-43, were found to have decreased rates of postabsorptive and insulin-stimulated glucose turnover that appear to result from decreased rates of glucose oxidation. In addition, G/Ghomozygote s were found to have reduced CAPN10 mRNA expression in their skeletal muscle. A decreased rate of insulin-mediated glucose turnover, or insulin resistance, is one mechanism by which the polymorphism in CAPN10 may increase susceptibility to type 2 diabetes mellitus in older person.”

Now that we have extracted the core component from the abstract, the document needs to be represented as a matrix format. Once we have this, there is still the problem of noise within the document. For example, the symbols in line 3 does not add any coherent meaning to the document, and therefore should be removed. Within the same document, there are symbols such as the forward slash (/) or hyphen (-). These are symbols (and others) that will not contribute to the word-document distribution. With the string module in python, these symbols and punctuations were removed with string.punctuation function. Eventually, we end up with a dataset of a similar format, ready to be further pre-processed and cleaned for modelling. The datasets and the Python codes used in this paper, to derive the results illustrated in the following sections, are publicly available at https://github.com/chongniki/Topic-Modelling_Diabetes.

2.3 | Dimensionality Reduction

When producing a word-document matrix, the matrix will end up having large dimensions, which ultimately hinders the computation time. In order to reduce the dimensions, we apply a method called word lemmatization. Word lemmatization is a linguistics method that replaces the word to its normalised form. Normalised form of a word would mean the original word, minus the suffixes. For example, a word 'singing' the normalised form would be 'sing'. There are other forms of this word as well – 'sung', 'singer', 'sang'. All of these words will be converted back to its normalised form in order to reduce dimensionality. Lemmatization reverses the process of grammatical suffixing, returning it to the verb it originated from [69].

The less popular Lemmatizer counterpart – Word stemmer, aims to achieve a similar outcome in a different manner. Applying word stemming to a document seems to be a more aggressive method, where the word is chopped by removing the suffix at the end of the word. Not to be confused with the lemmatizing method, where the word is transformed into its original state. For example, the same word 'singing' would be turned into 'sing'. Seemingly doing the same thing, the word stemmer falls when it comes to words such as 'matrices' and 'matrix' [70]. The stemmer cannot recognise the difference between these two words, thus leaving it as is. For this reason, the word lemmatizer is used to reduce the dimensionality of our matrix. To achieve this, the python NLTK has a preprogramed WordNetLemmatizer module, lemmatizing the verbs, adjectives, nouns, and adverbs.

Furthering the process of reducing dimensionality, stop words are words that exist in documents more frequently than the target words. It is theorised that words that appear the most frequent in a document contribute the least to the meaning of the entire text [71]. Words such as 'a' 'the' 'is' are examples of stop words that do not add much value to the context but is necessary to maintain the grammatical structure of a sentence. This was achieved by using the string module in Python, where it has a pre-defined set of stop words in the English language to reference to.

2.4 | Term Frequency Inverse Document Frequency (TF-IDF)

The first step in identifying the top words within a topic is to perform term frequency-inverse document frequency on the entire corpus. The intuition behind TF-IDF is to determine the frequency of each word in the document. The basis of this is on the assumption that words that have a higher relevance within the document tend to appear less frequently compared to those who appear more frequent [72]. In section 2.2, we looked into removing the stop words within a document. This process merely removes the 'padding' in a document. There exist words not included in the stop words set, that might not hold much significance in the topic evaluation. Thus, the words that appear more frequent within a document are lowered in terms of significance under TF-IDF. The Gensim package by [73] provides topic modelling modules in Python for developers. This includes the TF-IDF model for corpus creation, which was used for this study.

2.5 | Evaluation Topic Coherence

In section 1.5.4, we discussed the process of determining topic coherence in the model. The coherence score determines the ideal number of topics for the given dataset. That said, the $\mathsf{C_v}$ measure was chosen to evaluate the dataset from PubTator. In order to do so, we iteratively measured the coherence scores of each of the models tested with a different number of topics ranging from 1 to 200. By iteratively doing so we will be able to obtain the coherence scores for each number of topics and select the largest one from the list. This brute force method was computationally expensive, but essential.

The $\mathsf{C_v}$ measure will range from $\mathsf{0} < \mathsf{x} < \mathsf{1}$ as the output. By plotting the scores, we can see that the model peaks at 4 topics, with a $\mathsf{C_v}$ coherence score of 0.53 (2 d.p.). This score under the $\mathsf{C_v}$ measure indicates that the coherence of topic is satisfactory.

2.6 | Model Evaluation

Following the pre-processing steps taken, the output of the model results in a list of topics, and within each topic is a set of the top 10 words within that topic. Alongside the words are the word probabilities in the topics, indicating its prevalence within the topic. Topics coherence plot is depicted as Figure 6, and the output list is as follows:

```
[(0, '0.024*"gene" + 0.020*"risk" + 0.020*"association" + 0.019*"diabetes" + 0.018*"snp" + 0.017*"study" + 0.016*"associate" + 0.015*"type" + 0.013*"variant" + 0.011*"population"'),
(1, '0.038*"mutation" + 0.034*"diabetes" + 0.031*"mody" + 0.024*"gene" + 0.019*"hnf" + 0.019*"onset" + 0.018*"patient" + 0.012*"young" + 0.012*"factor" + 0.011*"family"'),
(2, '0.026*"cell" + 0.024*"insulin" + 0.019*"glucose" + 0.016*"expression" + 0.013*"gene" + 0.012*"beta" + 0.012*"diabetes" + 0.011*"secretion" + 0.009*"islet" + 0.009*"type"'),
(3, '0.038*"diabetes" + 0.024*"type" + 0.016*"study" + 0.016*"risk" + 0.013*"patient" + 0.012*"association" + 0.012*"associate" + 0.012*"variant" + 0.012*"gene" + 0.012*"polymorphism"')]
```

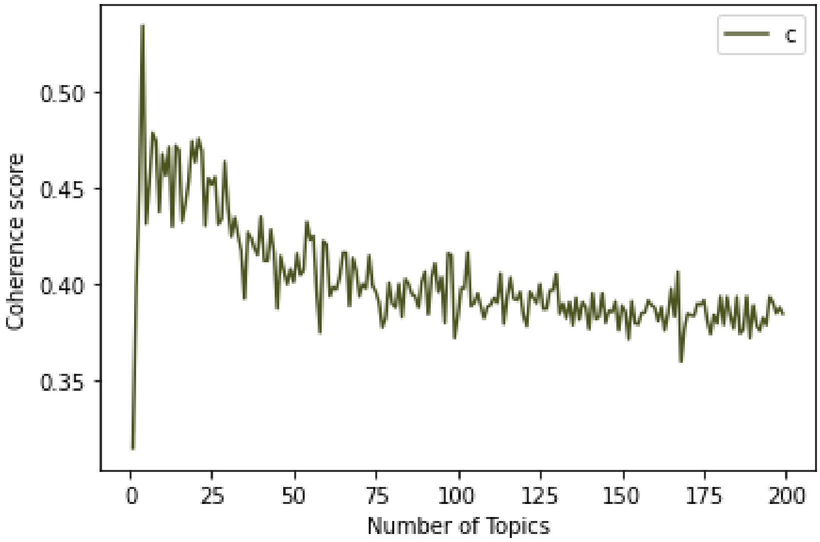


FIGURE 6 Topic Coherence plot for combined gene texts.

With the Gensim package, module PyLDAvis produces a more visually interactive figure for ease of interpretation. In Figure 7, we can see the 4 topics plotted on an inter-topic distance map, representing corpus probabilities. On the right, we see the top 30 words within the corpus after TF-IDF and LDA modelling. The circles represent the identified topics within the corpus, and their size represent the topic presence within the corpus. The distance between the circles demonstrates the similarity between each topic. The meanings of each topic will be discussed in the next section.

3 | RESULTS

3.1 | Critical Evaluation of Topic 1

In this section, we shall critically evaluate topic 1 as its corresponding LDA figure is illustrated in Figure 8. Overall, topic 1 has the highest prevalence within the corpus, accounting for 33.9% of the tokens within the identified corpus.

3.1.1 | Mutation

The reason a body can develop diabetes is due to the insulin gene mutation [74]. Glaser stated that genetic mutation can explain 60% of the cause of diabetes on the insulin gene [75].

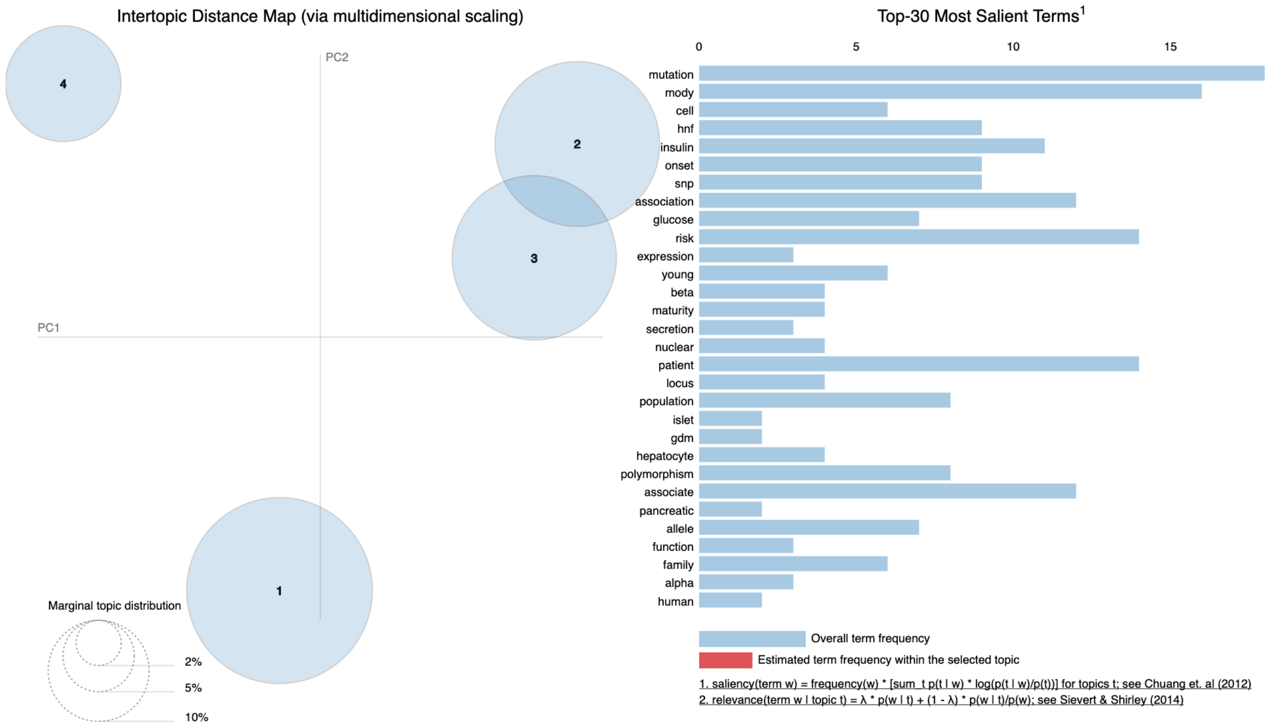


FIGURE 7 PyLDAvis plot for topics identified by LDA.

3.1.2 | MODY & Monogenic

Maturity-Onset Diabetes of the Young is a rare form of Diabetes that accounts for 1-2% of the diabetes population within the [76]. The top genetic cause of MODY has been identified to be the HNF1A, which has been found in 30-65% of MODY patients [77]. MODY falls under the category of Monogenic diabetes, which is a category of diabetes that is the result of genetic mutations, commonly found in the younger population [78].

3.1.3 | HNF & Hepatocyte

The hepatocyte nuclear factor 1 alpha is a common mutation of MODY. There have been 5 different mutations of this gene associated with Diabetes [79, 80].

3.1.4 | GCK (Glucokinase)

The role that GCK plays is associated with the recognition of glucose levels in the body [81]. It is responsible for the responsiveness of insulin production, depending on the glucose levels.

3.2 | Critical Evaluation of Topic 2

In this section, we shall critically evaluate topic 2 as its corresponding LDA figure is illustrated in Figure 9. Topic 2 accounts for 26.6% of the corpus.

3.2.1 | Single Nucleotide Polymorphisms (SNP)

This is a form of genetic variation found in diabetes patients. It affects the DNA nucleotides and have been found to affect non-obese diabetic patients [82].

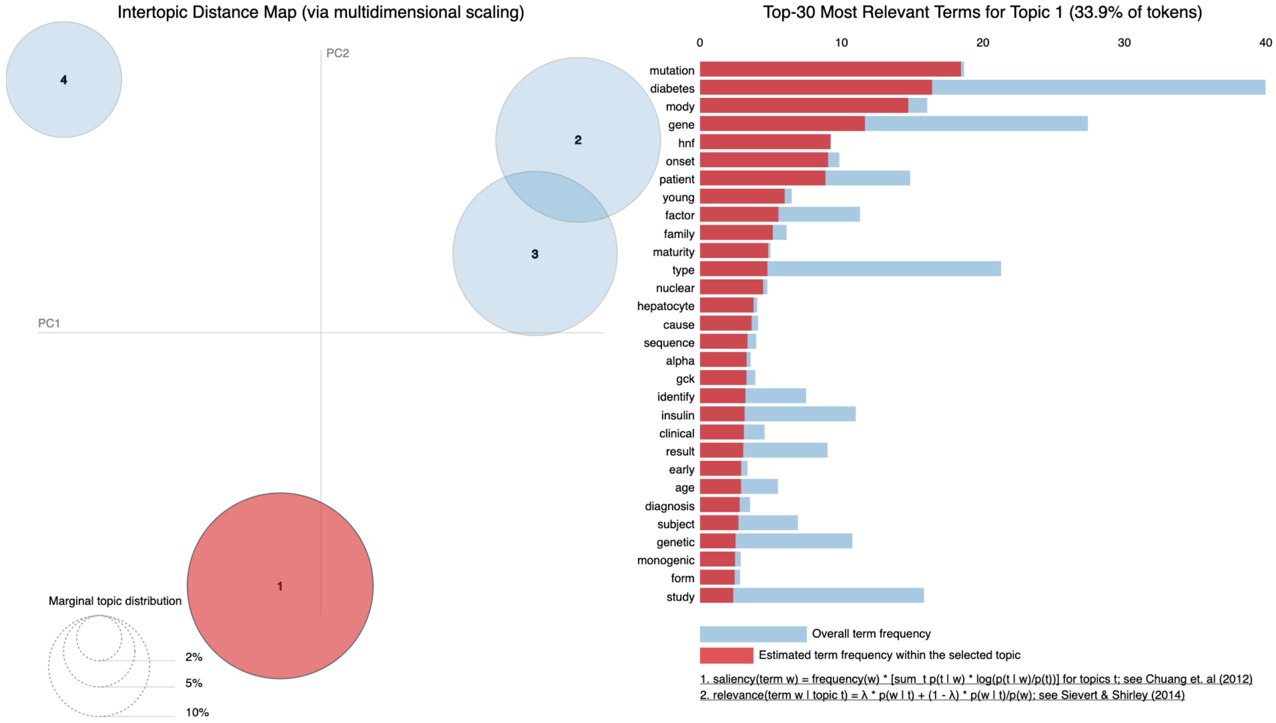


FIGURE 8 PyLDAvis plot for topic 1 identified by LDA.

3.2.2 | Allele

This term refers to the various forms of genes. Within a diabetes context, it is often used to describe an association between a particular gene variation and a characteristic. For example, an association between the risk TCF7L2 allele being inherited maternally and an increased risk of diabetes have been identified by [83, 84].

3.2.3 | Gestational Diabetes Mellitus (GDM)

[85] describes GDM as a form of diabetes that inhibits the tolerance of carbohydrates in pregnant women. Their study goes on to demonstrate that the TCF7L2 gene has been found to put pregnant women more at risk in developing the disease.

In relation to the previous term (GDM), the word ‘woman’ was picked up due to the fact that GDM is found in pregnant women [86].

3.3 | Critical Evaluation of Topic 3

The terms within Topic 3 appear to have similarities with Topic 2, as shown in Figure 10. Topic 3 accounts for 26.4% of the corpus.There are recurring words between the 2 topics. This association is demonstrated by the distance between both circles, indicating that the terms within each topic are rather similar to one another.

On of the most important term as illustrated in Figure 10 is Polymorphisms which refers to the genetic variation that occurs within a particular population sample ([87]).

Topics 2 and 3 can be analysed further by presenting them in terms of the first 10 terms with the highest frequencies:

Topic 2 - (0, ‘0.024*"gene" + 0.020*"risk" + 0.020*"association" + 0.019*"diabetes" + 0.018*"snp" + 0.017*"study" + 0.016*"associate" + 0.015*"type" + 0.013*"variant" + 0.011*"population"’)
Topic 3 - (3, ‘0.038*"diabetes" + 0.024*"type" + 0.016*"study" + 0.016*"risk" + 0.013*"patient" + 0.012*"association" + 0.012*"associate" + 0.012*"variant" + 0.012*"gene" + 0.012*"polymorphism"’)

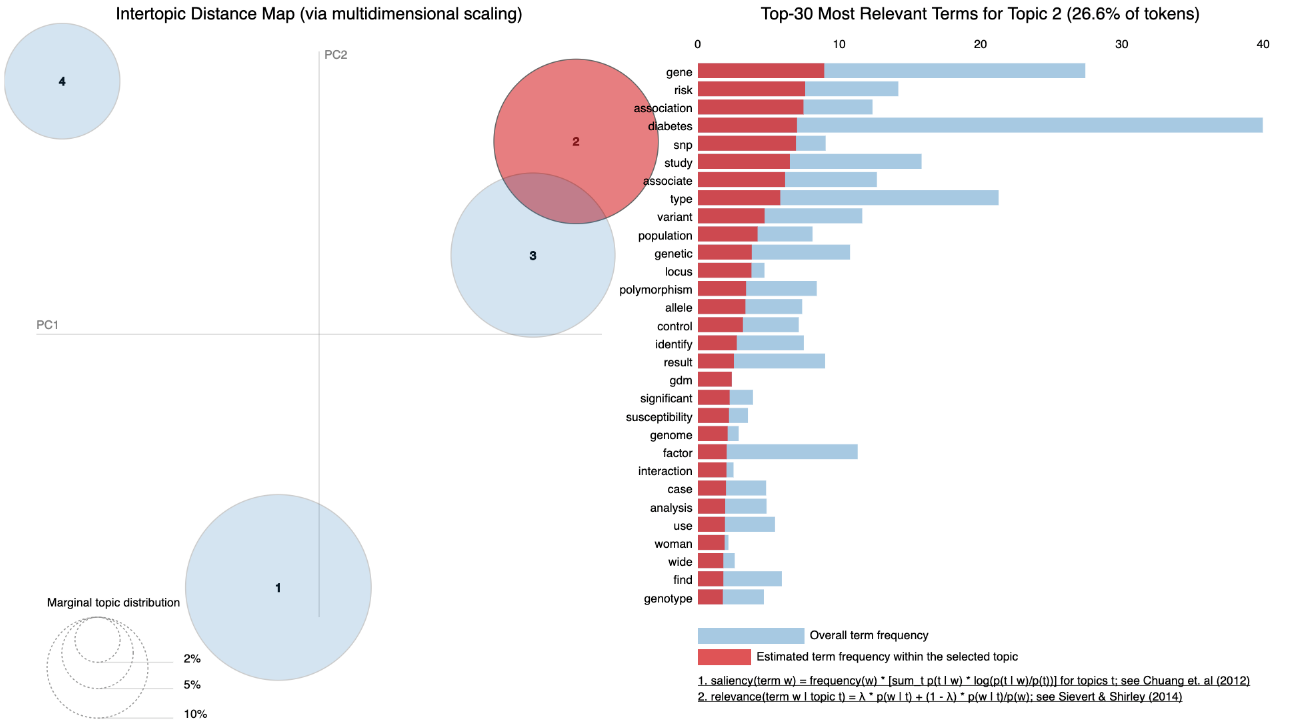


FIGURE 9 PyLDAvis plot for topic 2 identified by LDA.

Both the top 10 terms within these 2 topics are nearly similar, and the main difference is the weight of each term within the topics. Perhaps this is where the weakness of lack of data shows. Both these topics are identified as separate topics (although slightly similar), however, lack of data and terms within the corpus made it difficult for the model to extract the key differences from both these topics, resulting in a nearly similar topic assignment.

3.4 | Critical Evaluation of Topic 4

Finally, topic 4 accounts for 13.1% of the corpus, and the frequencies of the associated terms for this topic, estimated using the LDA-based topic modelling, are illustrated in Figure 11. We now briefly examine four determining terms to this topic and their relationships to type 2 diabetes: pancreatic, islet, beta, "wnt" (signalling pathways).

The role of pancreas in the body is to produce insulin glucose regulation. Any constraint to do so will result in the development of Diabetes [88].

Beta (β) cells are unique cells in the pancreas that produce, store and release the hormone insulin. Located in the area of the pancreas know as the islets of Langerhans (the organ's endocrine structures), they are one of at least five different types of islet cells that produce and secrete hormones directly into the bloodstream. Islet is simply a tissue that is particularly distinct from the surrounding ones. When the pancreas fails to produce sufficient insulin due to β-cell dysfunction, this cell dysfunction is seen in both T1D and T2D [89]. In T1D, the β dysfunction is often solved through islet transplantation.

The role of "wnt" signalling pathways is related to cellular developmental process among regular cells [90]. A defect in this, results in a decrease of insulin signalling between cells, and consequently reduced insulin secretion caused by the pancreas [91].

4 | DISCUSSION

In this study, we outlined the potential for the application of Topic Modelling (namely LDA) in Precision Medicine for Diabetes. The complexity of Diabetes requires for treatments to be tailored for each patient. The need for Precision Medicine has been frequently highlighted over the years since it was first introduced. Considering the results, it can be deduced that the statistical analyses of textual

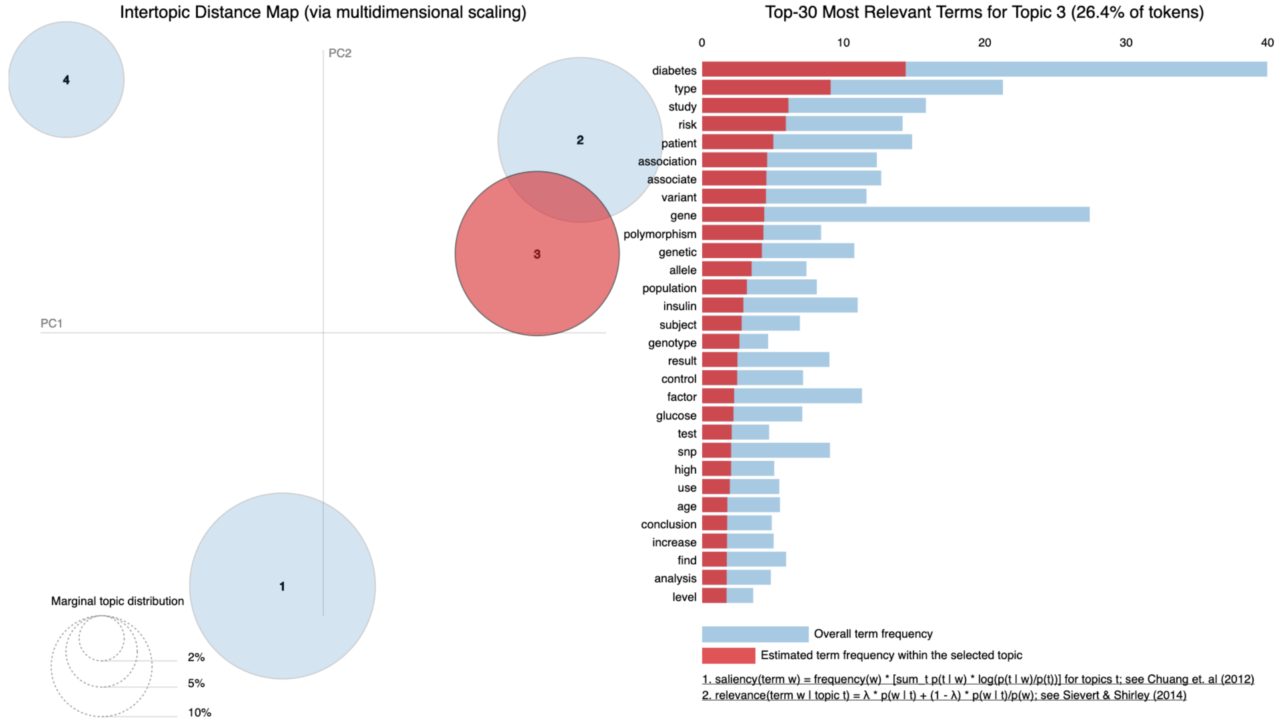


FIGURE 10 PyLDAvis plot for topic 3 identified by LDA.

data (journal abstracts) is capable of producing the key terms found in the disease (diabetes). The intended outcome of the study was to identify hidden associations between particular genes and the diabetes and based on the results, it has subjectively done so. With the 3 genes that were modelled, HNF1A, CAPN10, and TCF7L2, 4 topics were extracted based on the topic coherence scores (C_v).

Upon further evaluation of the topic terms, the model identified biological phenomenon that occurs among these 3 risk genes through the identified topics. It appears that topic 1 is highly related to the form of diabetes known as Maturity-Onset Diabetes of the Young. This form of diabetes has been identified to affect the HNF1A risk gene. Topics 2 and 3 appears to outline a form of genetic mutation that affects the DNA of diabetic patients. Within topic 2, gestational diabetes mellitus was identified, alongside this was the term “woman”. This is because gestational diabetes mellitus is found in 7% of pregnant women, that have a variant of the TCF7L2 risk gene [92]. Topic 4 highlights the pancreas’ role in the disease. Terms such as “beta”, “wnt”, and “islet” were identified, indicating the topic connotation. These are terms that relate to the pancreatic functions when a patient is diagnosed with diabetes.

These topics produced an adequate result given the dataset and tools used. Upon further inspection of these topics, there were recurring terms that were abbreviated and some not. This is because the lemmatizer is not capable of recognizing abbreviated terms due to the variations of abbreviations a term might have. Using the lemmatizer as a reference point, creating a dictionary of common terms and words for a particular subject such as diabetes may be able to improve the model by reducing the recurring words within the document. A downfall that I have identified is the lack of domain knowledge the topic model has when only statistically evaluating the corpus.

While the model did not manage to identify potential treatments and solutions to the disease, it was able to identify the topics commonly associated with a particular gene, which indicates the potential for the method to be further developed for the purpose of precision healthcare. The possible use cases for this method in diabetes research is identifying associations between a particular phenotype or gene, with a successful treatment outcome. This could possibly be further developed for the use of recommender systems, whereby a clinician can use the software as a guide to the diagnosis or treatment to a certain individual that falls within the same category or topic. [93] illustrates a similar idea with the usage of topic models for an e-commerce buying and selling platform. Based on the descriptions provided by the seller, topic modelling is applied to the descriptions. A customer would then be matched with the seller based on the topic similarity of the description and the customer’s needs. In the context of diabetes treatment, a similar topic-based recommender system can be applied on existing patient case study data, which contain descriptions of the patient’s reaction to a particular drug. This method can be used to aid the process of prescribing a potential treatment or drug to the patient, based on their genetic and phenotype profile.

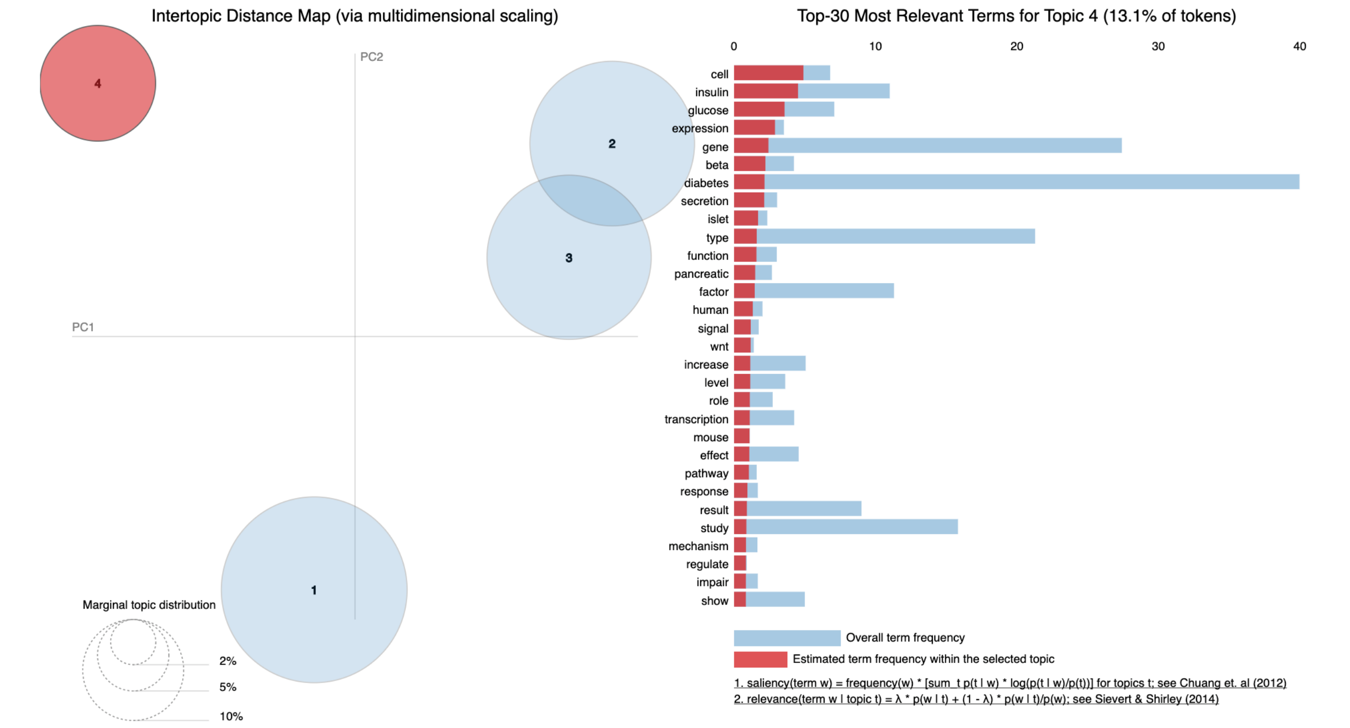


FIGURE 11 PyLDAvis plot for topic 4 identified by LDA.

Moving forward, the outcome of topic models can be further developed with relevant case study data. Case study data have the potential to identify traits and characteristics of a population and training the topic model on such data could derive valuable insight from the model. An area that was brought to light during this study is the use of topic modelling for big data. Naturally, the healthcare industry would be a huge source of big data and there needs to be means to address the usage of it with topic modelling methods. Even with a dataset of 1426 documents, the algorithm took 4 hours to run on a local machine.

A potential area to also be researched in conjunction with this is the combination of linguistic topic models with statistical topic models for the stated purpose. Using a set of pre-defined medical terms or abbreviations may aid the statistical topic model in finding more coherent topics than using just statistical methods. This is because the medical field uses terms and words that may not be common to the regular layman, let alone statistical topic models, which is worthwhile to look into.

5 | CONCLUSIONS

Based on the outcome of the study, it is theorised that there is potential for Topic Modelling methods to be applied for precision medicine. In consideration of the topic model results and its limitations, the topic model is able to statistically extract significant topics that occur within the dataset of 3 genes. That being said, it is worth noting that within an industry as paramount as healthcare, developing systems and solutions around natural language processing methods has its evident limitations. Statistically evaluating text merely identifies the highest (or lowest) probably occurrences within the document, however, there is an absence of domain specific knowledge that would substantially improve the outcome of the model. With the aforementioned dictionary, the model would probably be able to reduce the number of recurring words, thus improving the algorithm as a whole.

ACKNOWLEDGMENTS

This research was supported by funding from Coventry University for Research Internship Programme titled “Probabilistic modelling for Behaviour Analysis of the social media data”.

Author contributions

Conceptualisation, AD.; methodology, AD.; investigation, NC, AD; writing—original draft preparation, NC, AD, AHF; writing—review and editing, NC, AD, AHF, NS; visualisation, NC, AD, AHF, NS; supervision, AD, NS; project administration, NC. All authors have read and agreed to the final version of the manuscript.

Financial disclosure

None reported.

REFER

[1] David M. Blei. “Probabilistic Topic Models”. In: Commun. ACM 55.4 (2012), pp. 77–84. issn: 0001-0782. doi: 10.1145/2133806.2133826. url: https://doi.org/10.1145/2133806.2133826.

[2] Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. “Modeling healthcare data using multiple-channel latent Dirichlet allocation”. In: Journal of biomedical informatics 60 (2016), pp. 210–223.

[3] Adnan Muhammad Shah et al. “What patients like or dislike in physicians: Analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach”. In: Information Processing & Management 58.3 (2021), p. 102516.

[4] Chat Room. “Topic Modelling”. In: algorithms 9.07 (2020), p. 34.

[5] Pooja Kherwa and Poonam Bansal. “Topic modeling: a comprehensive review”. In: EAI Endorsed transactions on scalable information systems 7.24 (2020).

[6] Hanqing Xue et al. “Review of drug repositioning approaches and resources”. In: International journal of biological sciences 14.10 (2018), p. 1232.

[7] Giup Jang et al. “PISTON: Predicting drug indications and side effects using topic modeling and natural language processing”. In: Journal of biomedical informatics 87 (2018), pp. 96–107.

[8] Victor Chang. “Data analytics and visualization for inspecting cancers and genes”. In: Multimedia tools and applications 77.14 (2018), pp. 17693–17707.

[9] Victor Chang. “Computational intelligence for medical imaging simulations”. In: Journal of medical systems 42.1 (2018), pp. 1–12.

[10] Omid Chatrabgoun, Amin Hosseinian-Far, and Alireza Daneshkhah. “Constructing gene regulatory networks from microarray data using non-Gaussian pair-copula Bayesian networks”. In: Journal of Bioinformatics and Computational Biology 18.04 (2020), p. 2050023.

[11] Omid Chatrabgoun et al. “Approximating non-Gaussian Bayesian networks using minimum information vine model with applications in financial modelling”. In: Journal of computational science 24 (2018), pp. 266–276.

[12] Jinyu Hu et al. “Gene-based collaborative filtering using recommender system”. In: Computers & Electrical Engineering 65 (2018), pp. 332–341.

[13] Juan Zhao et al. “Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein (a)(LPA)”. In: PloS one 14.2 (2019), e0212112.

[14] Bahar Farahani et al. “Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare”. In: Future Generation Computer Systems 78 (2018), pp. 659–676.

[15] Maryam Farsi et al. Digital Twin Technologies and Smart Cities. Springer, 2020.

[16] Amin Hosseinian-Far, Muthu Ramachandran, and Charlotte Lilly Slack. “Emerging trends in cloud computing, big data, fog computing, IoT and smart living”. In: Technology for Smart Futures. Springer, 2018, pp. 29–40.

[17] Luca Catarinucci et al. “An IoT-aware architecture for smart healthcare systems”. In: IEEE internet of things journal 2.6 (2015), pp. 515–526.

[18] Sohail Jabbar et al. “Semantic interoperability in heterogeneous IoT infrastructure for healthcare”. In: Wireless Communications and Mobile Computing 2017 (2017).

[19] Liane Margarida Rockenbach Tarouco et al. “Internet of Things in healthcare: Interoperatibility and security issues”. In: 2012 IEEE international conference on communications (ICC). IEEE. 2012, pp. 6121–6125.

[20] Byung Mun Lee and Jinsong Ouyang. “Intelligent healthcare service by using collaborations between IoT personal health devices”. In: International Journal of Bio-Science and Bio-Technology 6.1 (2014), pp. 155–164.

[21] Sankalp Deshkar, RA Thanseeh, and Varun G Menon. “A review on IoT based m-Health systems for diabetes”. In: International Journal of Computer Science and Telecommunications 8.1 (2017), pp. 13–18.

[22] Ruhani Ab Rahman et al. “IoT-based personal health care monitoring device for diabetic patients”. In: 2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). IEEE. 2017, pp. 168–173.

[23] Munish Bhatia et al. “Internet of things-inspired healthcare system for urine-based diabetes prediction”. In: Artificial Intelligence in Medicine 107 (2020), p. 101913.

[24] Robert SH Istepanian et al. “The potential of Internet of m-health Things “m-IoT” for non-invasive glucose level sensing”. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE. 2011, pp. 5264–5266.

[25] Yang Yang et al. “Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system”. In: Information Sciences 479 (2019), pp. 567–592.

[26] Victor Chang et al. “Ethical problems of smart wearable devices”. In: 4th International Conference on Complexity, Future Information Systems and Risk. SciTePress. 2019, pp. 121–129.

[27] Jonathan H Chen et al. “Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets”. In: Journal of the American Medical Informatics Association 24.3 (2017), pp. 472–480.

[28] Cao Xiao et al. “Adverse drug reaction prediction with symbolic latent dirichlet allocation”. In: Proceedings of the thirty-first AAAI conference on artificial intelligence. 2017.

[29] Dimitris Bertsimas et al. “Personalized diabetes management using electronic medical records”. In: Diabetes care 40.2 (2017), pp. 210–217.

[30] C Zhang and R Zhang. “More effective glycaemic control by metformin in African Americans than in Whites in the prediabetic population”. In: Diabetes & metabolism 41.2 (2015), pp. 173–175.

[31] Lutz Heinemann et al. “Benefit of digital tools used for integrated personalized diabetes management: Results from the PDM-ProValue study program”. In: Journal of Diabetes Science and Technology 14.2 (2020), pp. 240–249.

[32] Oliver Baclic et al. “Challenges and opportunities for public health made possible by advances in natural language processing”. In: Canada Communicable Disease Report 46.6 (2020), pp. 161–168.

[33] J-D Kim et al. “GENIA corpus—a semantically annotated corpus for bio-textmining”. In: Bioinformatics 19.suppl_1 (2003), pp. i180–i182.

[34] Bach Xuan Tran et al. “Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study”. In: Journal of medical Internet research 21.11 (2019), e15511.

[35] Jeffrey Braithwaite et al. “Built to last? The sustainability of healthcare system improvements, programmes and interventions: a systematic integrative review”. In: BMJ open 10.6 (2020), e036453.

[36] Philip Resnik et al. “Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language”. In: Perspectives in Health Information Management (2008).

[37] E Ogbuju and GN Obunadike. “Information Extraction from Electronic Medical Records using Natural Language Processing Techniques”. In: Journal of Applied Sciences and Environmental Management 24.6 (2020), pp. 1027–1033.

[38] Novartis. Voluntary data sharing - Novartis: Clinical trials. Available from: https://www.novartisclinicaltrials.com/TrialConnectWeb/voluntarydatasharing.nov. 2016.

[39] Guifeng Xu et al. “Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study”. In: BMJ 362 (2018). issn: 0959-8138. doi: 10.1136/bmj.k1497. eprint: https://www.bmj.com/content/362/bmj.k1497.full.pdf. url: https://www.bmj.com/content/362/bmj.k1497.

[40] Amit Gupta, Tapan Behl, and Monika Sachdeva. “Key milestones in the diabetes research: A comprehensive update”. In: Obesity Medicine 17 (2020), p. 100183.

[41] Salvatore Carbone et al. “Glucose-lowering therapies for cardiovascular risk reduction in type 2 diabetes mellitus: state-of-the-art review”. In: Mayo Clinic Proceedings. Vol. 93. 11. Elsevier. 2018, pp. 1629–1647.

[42] Michael Mambiya et al. “The Play of Genes and Non-genetic Factors on Type 2 Diabetes”. In: Frontiers in Public Health 7 (2019), p. 349. issn: 2296-2565. doi: 10.3389/fpubh.2019.00349. url: https://www.frontiersin.org/article/10.3389/fpubh.2019.00349.

[43] Maria J Redondo and Patrick Concannon. “Genetics of type 1 diabetes comes of age”. In: Diabetes Care 43.1 (2020), pp. 16–18.

[44] Omar Ali. “Genetics of type 2 diabetes”. In: World journal of diabetes 4.4 (2013), p. 114.

[45] Brigitte I. Frohnert et al. “Predictive Modeling of Type 1 Diabetes Stages Using Disparate Data Sources”. In: Diabetes 69.2 (2020), pp. 238–248. issn: 0012-1797. doi: 10.2337/db18-1263. eprint: https://diabetes.diabetesjournals.org/content/69/2/238.full.pdf. url: https://diabetes.diabetesjournals.org/content/69/2/238.

[46] Savitha Subramanian and Irl B Hirsch. “Personalized diabetes management: moving from algorithmic to individualized therapy”. In: Diabetes Spectrum 27.2 (2014), pp. 87–91.

[47] James M Keller, Michael R Gray, and James A Givens. “A fuzzy k-nearest neighbor algorithm”. In: IEEE transactions on systems, man, and cybernetics 4 (1985), pp. 580–585.

[48] Giang Thu Vu et al. “Modeling the research landscapes of artificial intelligence applications in diabetes (GAPRESEARCH)”. In: International journal of environmental research and public health 17.6 (2020), p. 1982.

[49] Dan M Roden et al. “Pharmacogenomics: the genetics of variable drug responses”. In: Circulation 123.15 (2011), pp. 1661–1670.

[50] Laura A Levit et al. “Implementing precision medicine in community-based oncology programs: Three models”. In: Journal of oncology practice 15.6 (2019), pp. 325–329.

[51] Column Series. “A novel clinical pathways approach to delivering regional-based clinical trials and patient care in a hybrid academic-community-based system”. In: J Clin Pathways 4.4 (2018), pp. 52–55.

[52] Nader Salari et al. “The prevalence of stress, anxiety and depression within front-line healthcare workers caring for COVID-19 patients: a systematic review and meta-regression”. In: Human resources for health 18.1 (2020), pp. 1–14.

[53] Nader Salari et al. “The prevalence of sleep disturbances among physicians and nurses facing the COVID-19 patients: a systematic review and meta-analysis”. In: Globalization and health 16.1 (2020), pp. 1–14.

[54] Abhinav Vepa et al. “Predicting mortality, duration of treatment, pulmonary embolism and required ceiling of ventilatory support for COVID-19 inpatients: A Machine-Learning Approach”. In: medRxiv (2021).

[55] Abhinav Vepa et al. “Using Machine Learning Algorithms to Develop a Clinical Decision-Making Tool for COVID-19 Inpatients”. In: International Journal of Environmental Research and Public Health 18.12 (2021), p. 6228.

[56] Alireza Daneshkhah et al. “Behavioural Analytics: A Preventative Means for the Future of Policing”. In: Policing in the Era of AI and Smart Societies. Springer, 2020, pp. 83–96.

[57] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. “An introduction to latent semantic analysis”. In: Discourse Processes 25.2-3 (1998), pp. 259–284. doi: 10.1080/01638539809545028.

[58] A. Navlani. Latent Semantic Analysis using Python. Available from: https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python. 2018.

[59] Said A Salloum, Rehan Khan, and Khaled Shaalan. “A survey of semantic analysis approaches”. In: Joint European-US Workshop on Applications of Invariance in Computer Vision. Springer. 2020, pp. 61–70.

[60] Jonathan Chang et al. “Reading Tea Leaves: How Humans Interpret Topic Models”. In: Advances in Neural Information Processing Systems. Ed. by Y. Bengio et al. Vol. 22. Curran Associates, Inc., 2009, pp. 288–296. url: https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

[61] David Mimno et al. “Optimizing semantic coherence in topic models”. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011, pp. 262–272.

[62] Michael Röder, Andreas Both, and Alexander Hinneburg. “Exploring the space of topic coherence measures”. In: Proceedings of the eighth ACM international conference on Web search and data mining. 2015, pp. 399–408.

[63] David Newman et al. “Automatic evaluation of topic coherence”. In: Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010, pp. 100–108.

[64] Jipeng Qiang et al. "Short text topic modeling techniques, applications, and performance: a survey". In: IEEE Transactions on Knowledge and Data Engineering (2020).

[65] Nikolaos Aletras and Mark Stevenson. "Evaluating Topic Coherence Using Distributional Semantics". In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers. Potsdam, Germany: Association for Computational Linguistics, Mar. 2013, pp. 13–22. url: https://aclanthology.org/W13-0102.

[66] Frances Rivera Dennis. "Improving Healthcare Services in the Rural Counties of the Palmetto State: A Factor Analysis Study". PhD thesis. Keiser University, 2020.

[67] Toni Cvitanic et al. "Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents". In: International Conference on Case-Based Reasoning. 2016.

[68] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA". In: Information Retrieval 14.2 (2011), pp. 178–203.

[69] Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations". In: PloS one 12.6 (2017), e0179488.

[70] Jinxi Xu and W Bruce Croft. "Corpus-based stemming using cooccurrence of word variants". In: ACM Transactions on Information Systems (TOIS) 16.1 (1998), pp. 61–81.

[71] Charu C Aggarwal and ChengXiang Zhai. "A survey of text clustering algorithms". In: Mining text data. Springer, 2012, pp. 77–128.

[72] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: Proceedings of the first instructional conference on machine learning. Vol. 242. New Jersey, USA. 2003, pp. 133–142.

[73] Radim Rehurek. "models. word2vec-Word2vec embeddings". In: Gensim: Topic Modelling for Humans (2019).

[74] Masahiro Nishi and Kishio Nanjo. "Insulin gene mutations and diabetes". In: Journal of diabetes investigation 2.2 (2011), pp. 92–100.

[75] Benjamin Glaser. "Insulin mutations in diabetes: the clinical spectrum". In: Diabetes 57.4 (2008), pp. 799–800.

[76] Diabetes UK. maturity onset diabetes of the young (mody). 2020. url: https://www.diabetes.org.uk/diabetes-the-basics/other-types-of-diabetes/mody (visited on 2020).

[77] Rochelle Naylor, D del Gaudio, et al. "Maturity-onset diabetes of the young overview". In: (2018).

[78] Andrew T Hattersley and Kashyap A Patel. "Precision diabetes: learning from monogenic diabetes". In: Diabetologia 60.5 (2017), pp. 769–777.

[79] Sian Ellard. "Hepatocyte nuclear factor 1 alpha (HNF-1α) mutations in maturity-onset diabetes of the young". In: Human mutation 16.5 (2000), pp. 377–385.

[80] Sumreen Begum. "Hepatic Nuclear Factor 1 Alpha (HNF-1α) In Human Physiology and Molecular Medicine". In: Current Molecular Pharmacology 13.1 (2020), pp. 50–56.

[81] Jakub Hulın et al. "CLINICAL IMPLICATIONS OF THE GLUCOKINASE IMPAIRED FUNCTION–GCK-MODY TODAY". In: Physiological Research 69 (2020), pp. 995–1011.

[82] Meijun Chen et al. "Three single nucleotide polymorphisms associated with type 2 diabetes mellitus in a Chinese population". In: Experimental and Therapeutic Medicine 13.1 (2017), pp. 121–126.

[83] Rachel M Freathy et al. "Type 2 diabetes TCF7L2 risk genotypes alter birth weight: a study of 24,053 individuals". In: The American Journal of Human Genetics 80.6 (2007), pp. 1150–1161.

[84] Robin N Beaumont et al. "Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics". In: Human molecular genetics 27.4 (2018), pp. 742–756.

[85] Nael Shaat and Leif Groop. "Genetics of gestational diabetes mellitus". In: Current medicinal chemistry 14.5 (2007), pp. 569–583.

[86] Camille E Powe et al. "Genetic Loci and Physiologic Pathways Involved in Gestational Diabetes Mellitus Implicated Through Clustering". In: Diabetes 70.1 (2020), pp. 268–281.

[87] Beska Z Witka et al. "Type 2 Diabetes-Associated Genetic Polymorphisms as Potential Disease Predictors". In: Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy 12 (2019), p. 2689.

[88] Martijn van de Bunt et al. "The miRNA profile of human pancreatic islets and beta-cells and relationship to type 2 diabetes pathogenesis". In: PloS one 8.1 (2013), e55272.

[89] Maria S Remedi and Christopher Emfinger. “Pancreatic β-cell identity in diabetes”. In: Diabetes, Obesity and Metabolism 18 (2016), pp. 110–116.

[90] Kevin Pruitt. Molecular and cellular changes in the cancer cell. Academic Press, 2016.

[91] Amy C Arnold and David Robertson. “Defective Wnt signaling: a potential contributor to cardiometabolic disease?” In: Diabetes 64.10 (2015), pp. 3342–3344.

[92] Serdar H Ural, John T Repke, et al. “Gestational diabetes mellitus”. In: Reviews in Obstetrics and Gynecology 1.3 (2008), p. 129.

[93] Konstantinos Christidis and Gregoris Mentzas. “A topic-based recommender system for electronic marketplace platforms”. In: Expert Systems with Applications 40.11 (2013), pp. 4370–4379.

□