National College of Ireland

Analysing the Evolution of FinTech Research Topics in Academia

MSc Research Project
MSc in FinTech

Mudit Agarwal
Student ID: x18108202

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mudit Agarwal		
Student ID:	X18108202		
Programme:	MSc in FinTech	Year:	2019
Module:	MSc Research Project		
Supervisor:	Victor Del Rosal		
Submission Due Date:	12/08/2019		
Project Title:	Analysing the Evolution of FinTech Research Topics in Academia		
Word Count:	6599	Page Count	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 16/09/2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysing the Evolution of FinTech Research Topics in Academia

Mudit Agarwal
Student ID: x18108202

Abstract

The FinTech or financial technology industry has burgeoned in the past decade and has emerged as one of the hottest industries in recent times thanks to a myriad of innovations and disruption. New technologies like artificial intelligence (AI), blockchain, regulatory technology (RegTech), robotic process automation (RPA), insurance technology (InsurTech), internet of things (IoT), cloud computing and many more have been implemented in various different industries in unprecedented ways thereby disrupting the status quo. KPMG indicates that investment in the FinTech industry has doubled from 2017 to 2018, hitting a record $111.8 billion. Owing to the enormous global investment in the FinTech sector and its rapid expansion, it becomes imperative to analyse the research topics directing the growth of the FinTech sector. This research analyses FinTech research done in the past decade to determine the hot topics and provide an overview of how these topics have changed over time. Abstracts of journals articles and conferences are taken and a Latent Dirichlet Allocation (LDA) model using Gibbs sampling is trained to extract topics from them using topic modelling. The trends of the extracted topics are also identified including the positive or negative linear trends as well as non-linear trends.

1 Introduction

Financial technology (FinTech) has been growing at a tremendous rate over the past decade which is evident from the massive investments made into this field all over the globe. Deloitte indicates that the number of FinTech start-ups entering the market during 2010-2012 doubled when compared to those entering the market in 2008-2010 (Deloitte, 2017). KPMG indicates that global investment in the financial technology sector has doubled from 2017 to 2018 and has hit a record high of $111.8 billion as seen in the figure below (KPMG, 2019).

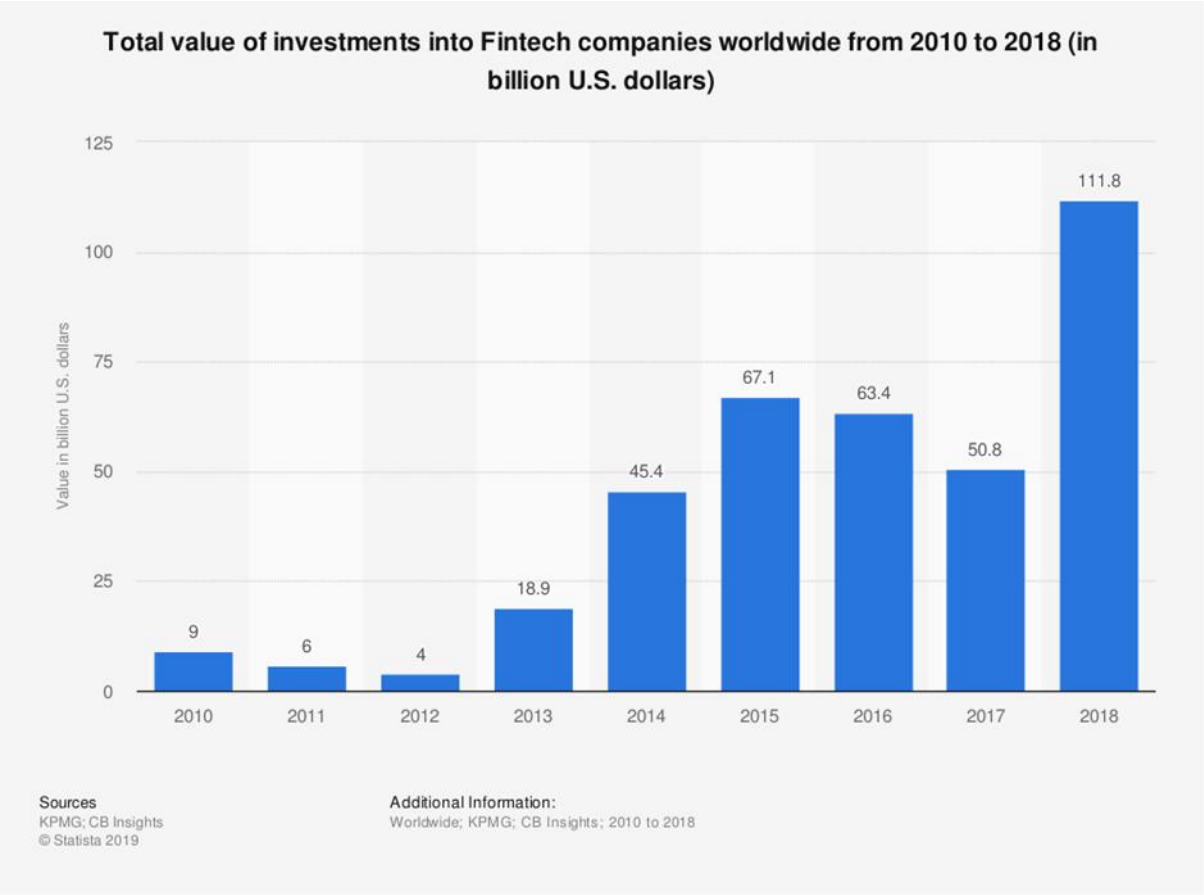


Figure 1: Global FinTech investment

Since the FinTech industry has clearly been burgeoning and immense capital is being injected into it, it is imperative that trends in the industry must be analysed so as to gain a deeper insight into where the industry is headed. This would give an understanding about the novel technologies being used to disrupt the status quo in the market which can be extremely valuable.

A plethora of study areas exist within the financial technology domain which are creating technology that is disrupting many different industries. Such exciting new areas of study are widely discussed and studied by students of FinTech. This enables them to excel academically while also narrow down their personal interests within FinTech. It also provides them with a deep understanding about which topics or areas of study are fuelling the rapid advancement of the FinTech industry. This research aims to provide a similar insight into various FinTech research topics that have been pushing the industry to the forefront over the past decade. Such information can be extremely valuable to multiple stakeholders and people for different reasons.

The leadership of FinTech companies and other organisations providing financial or FinTech-realted services ought to be very well versed with the different emerging technologies and their evolution so as to identify the ones that are projected to disrupt the status quo in the industry. This can be greatly facilitated by research such as this one as trends in academic research are a great indicator of the evolution of emerging technologies. This could potentially allow the company to focus their research and development (R&D) efforts in studying and implementing the identified emerging technologies ahead of the

competition thereby providing unique competitive advantage to the organisation. The lack of this foresight, however, could prove to be a massive missed opportunity costing organisations millions in revenue. This is evident from the way Amazon have become the world leaders in providing cloud services by being one of the first companies to innovate in that area of study.

Besides upper management/leadership of FinTech organisations and researchers/students studying financial technology, this research will also be useful to governments and policymakers as it will allow them to identify upcoming technologies ahead of time. The government and policymakers could then study and understand the emerging technologies and plan the regulations and laws needed to efficiently regulate the new technologies. By identifying the topics and technologies in advance, all this can be done before the technologies are actually put into widespread use thereby allowing smooth rolling out of the technologies and their adoption without any hesitation. Bad planning before rolling out novel technologies can result in negative perception of the technology which pulls back its adoption by years.

Text mining as an area of study or concept, was first created by Feldman & Dagan (1995) in 1995. Starting then, the field of text mining has seen massive developments resulting in a myriad of techniques that can be used for text mining and feature extraction. These include Naïve Bayes classification, K-Means Clustering and several more (Wang, et al., 2000). These techniques were quite efficient in analysing portions of text but did not capture the relationships between words and documents. This created a necessity for topic modelling. It was derived in 1990 from Latent Semantic Indexing (LSI) that was based on linear algebra (Deerwester, et al., 1990). Latent Semantic Indexing involves creating vector representations of the text being analysed which in turn are used to identify relationships between words. Hofmann (1999) gave rise to Probability Latent Semantic Analysis (PLSA) in 1999 when he proposed the use of probability to circumvent the need for complex calculations. Latent Dirichlet Allocation (LDA) was formed when the PLSA model was improved by including Dirichlet distribution in Probability Latent Semantic Analysis. In LDA, the text documents are regarded as massive mixtures of latent topics and the relationships among these topics are identified using Bayesian probabilistic distributions over words in the text documents. Latent Dirichlet Allocation is most suitable for summarising a large corpus of disparate unstructured text data sources and this is why it is one of the most extensively used techniques for topic modelling.

Figure 2 (Lee, et al., 2018) shows the LDA model and its working visually as seen below:

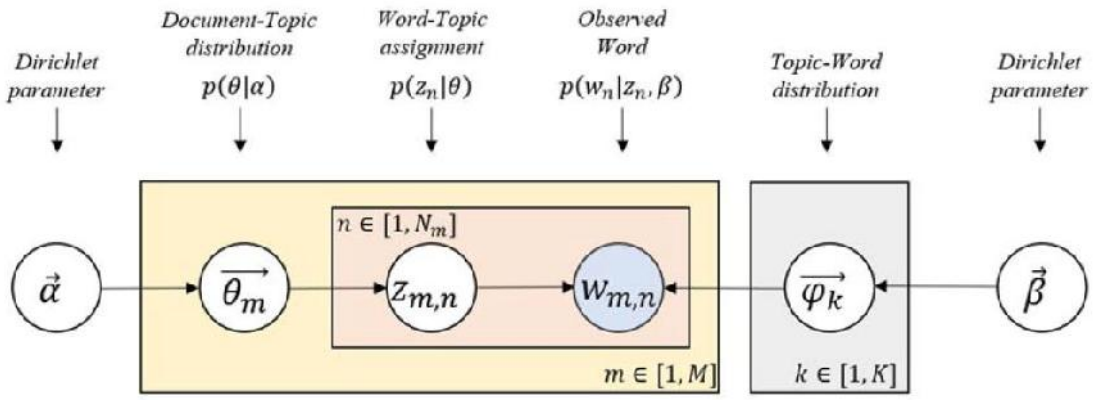


Figure 2: Graphical model of LDA

This paper is organised as follows: Section 2 discusses the related research done on topic modelling while Section 3 elucidates the research methodology that was followed for the course of this research. Section 4 describes the design specification and implementation. Section 5 provides the evaluation of the model and the paper is finally concluded in Section 6 which also lays out the future work.

2 Related Work

Bittermann & Fischer (2018) analyse academic research in psychology using Latent Dirichlet Allocation and show its advantages over the classification-based scientometric approaches. They use 314,573 psychology publications published from 1980 to 2016, comprising of articles, book chapters, reports and dissertations extracted from the PSYNDEX database for psychology academic literature. Latent Dirichlet Allocation using Gibbs sampling is employed for topic modelling and non-linear trends are identified using multilayer perceptrons (MLPs) with two hidden layers. The average topic probabilities as a non-linear function of the publication year are calculated by the multilayer perceptrons. Moreover, the optimal number of topics to be determined from the dataset was chosen as 500 by analysing the values of log likelihood for different models. Both linear and non-linear trends were determined and 128 topics were observed to display an upward linear trend while 135 portrayed a downward linear trend. The topics portraying the upward trend included topics like human migration, visual attention, online therapy, neuropsychology, genetics and traumatisation among several others. These were regarded as the hot topics in the field of psychology due to their recent upward trend. This method has advantages over classification-based techniques as this model provides more specific information on the topics. The research notes and iterates that emerging techniques like polylingual topic modelling, dynamic topic modelling and multilingual probabilistic topic modelling can provide improved results in the future. This research outlined in this paper will also use LDA with Gibbs sampling for topic modelling and log likelihoods will be one of the metrics used to determine the optimal number of topics to be extracted from the dataset.

Similarly, Amado et al. (2018) delve into big data in marketing and provide insights using topic modelling and text mining. Author affiliations are used to identify major contributors as well. Relevant domain sciences and non-technical terms in several areas of marketing are used to extract 1560 articles and journals from ScienceDirect published between 2010 and 2015. The subjects were analysed to form five representative dictionaries for geography, products, sectors and big data & marketing. The geographical dictionary determines the country contingent on the author's affiliations, the products dictionary determines products and services under question, the sectors' dictionary identifies the various sectors of the economy being considered and finally the big data & marketing dictionary determines the terms used. Topics are extracted from documents using topic modelling implemented by LDA. USA and China are observed to be the largest major contributors which is expected due to their large populations. India, Australia and Spain are also identified as major contributors providing significant contributions to academic research in the field. The tremendous growth in big data and marketing is evident upon observing that the number of academic publications in this field has doubled year on year. It is observed that Europe, Asia and North America are major contributors in the space. Furthermore, it is noted that healthcare and energy garner 50% of the attention given to consumer goods in North America.

Chanda & Das (2018) employ a graph-based clustering technique to cluster documents together and identify their topics using the importance factor of documents. A corpus of 18,834 documents is extracted from newsgroup20 consisting of twenty topics. Next, stop words from this dataset are deleted and words are turned to their root form through the process of stemming. Following this, the root words are converted to vectors using inverse document frequency (IDF) and term frequency (TF). These vectors are in turn used as the input for the clustering algorithm. The weightage of documents is determined by their importance factor and the node with the highest importance factor becomes a cluster centre. When adjacent nodes have a similarity weight that exceeds the threshold value, they are included in the cluster and the new whole is then considered a node. This process is performed repeatedly until the similarity weights of all neighbouring nodes is lesser than the threshold value. Abulaish & Fazil (2018) analyse tweets from the social media platform Twitter and show variations in user behaviour in terms of the topics discussed they discuss on the social media platform. A word embedding based approach is used for this research.

Text data like comments from StackOverflow are analysed by Johri & Bansal (2018) in order to identify hot topics being discussed and technology trends in the technology and computer science field. From the dataset comprising of comment data and other user generated content, stop words, numbers, code and urls are eliminated following which the process of stemming is performed in order to convert the words to their root form. This step is similar to the research discussed previously (Chanda & Das, 2018). Next, Latent Dirichlet Allocation is performed on the transformed data to extract 40 topics from the dataset. The top terms are identified in each topic and these are used to manually label each topic. Hot topics and most impactful topics are identified and their trend in analysed. From a total of 40 topics, 16 topics are found to be having an upward linear trend, 15 having a downward linear trend and 9 having a steady trend. Competing programming languages are identified and the most popular one among them are also identified. It is observed that website design/CSS is the

most impactful while mobile app development and data analysis/visualisation are the hottest topics. This research provides a deep insight about which technologies and languages are most widely used, their relationships with one another and also their impact on their respective industries. The research outlined in this paper seeks to provide similar insights into the prevailing trends in the FinTech domain.

A two-phased technique for topic modelling is used by Wai & Aung (2018) combining Latent Dirichlet Allocation with pattern mining techniques that provides more specific and improved results as opposed to the results obtained by using LDA alone. Two datasets of paper abstract data containing 548 and 129000 abstracts are used as datasets for this research. Stop words are eliminated and stemming is performed similar to the research discussed earlier (Johri & Bansal, 2018). Next, LDA is applied and twenty topics are extracted from each dataset. Topical transaction datasets are then created to optimise topic representations which in turn is used to create frequent item sets. Information entropy is used to measure performance. The frequent item sets based model is observed to have lower entropy compared to the baseline model thereby resulting in more specific topic representations.

Abuhay et al. (2018) extract research topics from journal papers published in the International Conference on Computational Science (ICCS) using non-negative matrix factorisation topic modelling. Autoregressive Integrated Moving Averages (ARIMA) is used to predict trends in research topics in the future. For this research, 5928 journal papers published between 2001 and 2007 were extracted from the ICCS and non-negative matrix factorisation is performed for topic modelling. The results obtained were further analysed using Change Point Analysis (CPA) and trend analysis. The time series is first stationarised before applying ARIMA for trend prediction. Model performance is measured using Root Mean Squared Error (RMSE) and the prediction performance is found to be fairly accurate. The research notes that multivariate time series prediction may further increase the accuracy of the model.

Liu et al. (2018) analyse topics being discussed in the online discussion forums of the "Introduction to Psychology" degree course at a Chinese university. Emotion driven topic modelling is performed on student comments and the emotions being reflected are obtained and classified into positive or negative emotions and confusion. Moreover, topics being discussed are determined for each emotion, which provides an insight about which topics are perceived to be strengths by students and which are causing trouble for students. This allows the identification of problem areas in which the students need help and this information can be very valuable to teachers and other stakeholders. Problem areas can be worked upon by the institution to improve the quality of education received by students and areas of strengths can be lessons that may be introduced into other course offerings thereby improving other courses based on these learnings. The research outlined in this paper does not require emotion based topic modelling but it is fascinating to look at its implications and the actionable results that can instantly be used to improve the services provided.

Wang & Yang (2018) perform a similar research by incorporating ratings and user sentiment in topic modelling on user product reviews. A Sentiment Topic Factorisation Model (STFM) is proposed for including user sentiment as features extracted by topic modelling through LDA, TopicMF, Hidden Factors and Hidden Topics (HFT) usually do not

take user sentiment into account. User preference is extracted from review sentiment using a lexicon method combined with a transform function while LDA is used for feature extraction. The item features and user preferences are then combined by the matrix factorisation model. 22 Amazon datasets obtained from McAuley & Leskovec (2013) are used for this research and the model performance is as good as the state of the art and even better for 13 datasets. The research notes that deep learning can provide even better accuracy in natural language processing (NLP) and can be utilised in the future to improve upon the model and achieve even better performance.

Vamshi et al. (2018) also combine sentiment analysis with aspect based opinion mining. Aspect based opinion mining is a sophisticated technique for feature extraction as well as determining user ratings for these features or aspects. This research uses tweets that include reviews from the social media platform Twitter as its primary dataset. This research, like several others discussed above, uses Latent Dirichlet Allocation for feature extraction and topic modelling. The sentiment analysis is done via Support Vector Machines (SVMs) which perform even better than Naïve Bayes. LDA identifies the topics or features being talked about from the unstructured text data and SVM identifies whether user sentiment towards the extracted features are positive or negative. The research also notes that it can suffer due to spam or incorrect opinions/reviews despite being a good technique for opinion mining and sentiment analysis. Fake news and spam opinions have been rising on social media platforms lately and an effective way to segregate authentic reviews from fake ones needs to be devised. Once fake opinions are eliminated from the dataset, the results can be trusted completely without any manufactured bias.

Laoh et al. (2018) analyse Indonesian song lyrics featuring the convoluted and nuanced language of Bahasa, to extract the topics of the songs using Latent Dirichlet Allocation. For the purpose of this research, 193 Indonesian songs are chosen from the top 200 Spotify list from January 2017 to January 2018. The lyrics of these songs are analysed to determine the topic or theme of the song. Multiple pre-processing steps are performed like tokenisation, case folding, stemming and filtering. These pre-processing steps are observed in research done by several others as discussed above. These steps prepare the data for applying LDA for topic modelling. The LDA model is evaluated using perplexity as a performance metric and it is also used to determine the optimal number of topics. Ten topics including “joy and party”, “sad and sorrow”, “love and romantic” and “nationalism” are identified and probabilities for each topic are calculated. The topic corresponding to the highest probability is chosen as the topic of that song. “Love and romantic” is observed to be the most popular/prevalent topic discussed most extensively. It should be noted that the size of the dataset is relatively small and a more holistic view can be provided if more data is used for topic extraction.

Hidayatullah et al. (2018) analyse football related tweets from the social media platform Twitter, taken from several Indonesian Twitter accounts that provide updates and commentary on football. Latent Dirichlet Allocation is used for topic extraction and these topics give an overview of which topics are being discussed most in the Indonesian football community. The hot topics obtained from LDA are then visualised through python libraries namely pyLDAvis and Gensim so as to depict the relationships between words and topics. Several topics are observed to be similar while others are found to be completely independent

of one another. Top 30 words appearing most frequently in the data are identified and an analysis of the top 5 terms portray that Manchester United, Real Madrid and Chelsea are the teams that are most widely discussed in the Indonesian football community. Analysis of the top hot topics discussed in the news reveal that “El Clasico”, “English Premiere League”, “pre-match analysis” and “live updates” are the topics.

The aforementioned research papers provide an overview of the disparate approaches being applied in the field of topic modelling. It is noted that most implementations of topic modelling or models that involve feature extraction are based on Latent Dirichlet Allocation, sometimes in conjunction with other techniques. The research outlined in this paper will apply topic modelling to obtain keen insights in the financial technology domain.

3 Research Methodology

This research follows the Knowledge Discovery in Databases (KDD) (Fayyad, et al., 1996) approach that comprises of various steps. These steps are shown below in Figure 2:

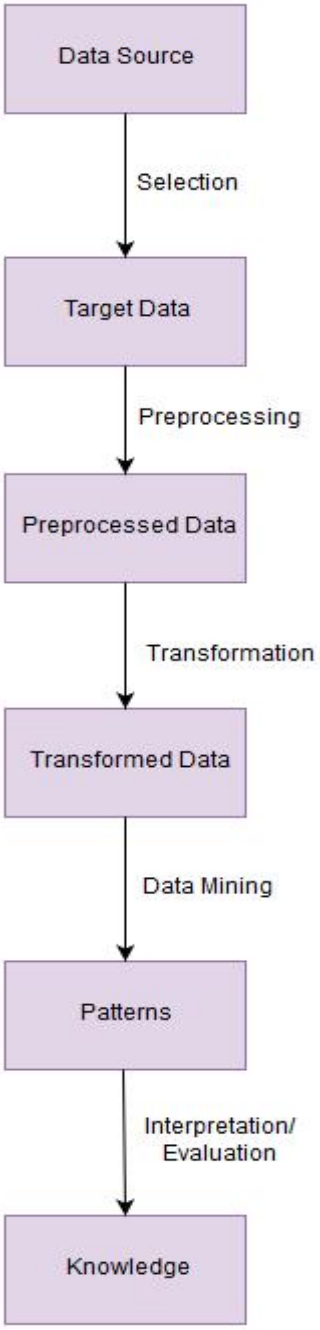


Figure 3: Knowledge Discovery in Databases (KDD)

A. Data Selection:

This is the first step of KDD and it includes selecting the most appropriate data from the data corpus or data source chosen for the analysis. The most relevant and appropriate data must be identified in terms of the knowledge discovery exercise undertaken and only this selected data is fetched and used for the purpose of the knowledge discovery exercise. This ensures that there is no superfluous data in the dataset used. This optimises the size and amount of data needed for analysis.

B. Data Cleaning & Pre-processing:

This next step in KDD is essential as it includes cleaning the data. Any missing data is either removed or dealt with using various techniques. Noise and outliers are also identified and dealt with either by deletion/removal or by using techniques to make the data ready for the next step, data transformation.

C. Data Transformation:

This step requires knowledge of the data analysis technique chosen to be performed on the dataset. The form of the data required to be fed into the data analysis method is identified and the purpose of this step is to transform the dataset into the identified form so as to achieve the best results. Data transformation methods are applied to the cleaned and pre-processed data to create the best representation of the most useful features relevant to the data analysis task undertaken.

D. Data Mining:

This is one of the most important steps in KDD as this step includes actually performing the analysis on the dataset using the data mining technique chosen for analysis. The technique chosen must be the most suitable technique for the kind of analysis to be done as the more appropriate the method, the better will be the results. The data mining method is then applied to the transformed data without any issues since the dataset is already in the correct form to be fed in to the method.

E. Interpretation/Evaluation:

This is the final step in KDD and it involves evaluating the analysis performed in the previous step. The performance of the model or technique used is measured using a plethora of metrics. Moreover, the results of the analysis are interpreted and explained so as to gain an idea about the impact or significance of the results. This understanding results in valuable knowledge that can be used to make further decisions.

For the purpose of this research, the following tasks are performed as part of the steps in the Knowledge Discovery in Databases cycle:

A. Data Selection:

- IEEE Xplore Digital Library (IEEE, 2019) is used as the primary data source for this research as it provides easy access to journal articles and conference proceedings in computer science and other related fields.
- FinTech journal articles and conference proceedings published from 2010 to 2019 are extracted from IEEE Xplore Digital Library in csv format. Various buzzwords/keywords are used as search strings to search for FinTech publications including blockchain, financial technology, fintech, blockchain, smart contracts, cryptocurrency, distributed ledger, financial services, initial coin offering, insurtech, regtech, DLT and ICO.

B. Data Cleaning & Pre-processing:

- The journal articles and conference proceedings belonging to different search strings are then combined/aggregated into a single csv file. There are separate columns for Abstract and Year of Publication which are the two most useful columns for this research.
- The aggregated dataset is then analysed and it is observed that many observations have numerical values and missing values for Abstract and Year of Publication attributes. These observations are removed from the dataset.
- It is observed that a total of 14,116 unique observations are obtained and these can be used for the purpose of this research.

C. Data Transformation:

- The dataset in csv form is ingested into R and all attributes are converted to string for ease of use as Abstract data should be in string format.
- Case folding is performed which involves converting the entire text to lowercase.
- Stop words like ‘a’, ‘an’ and ‘the’, which have no intrinsic meaning and therefore are irrelevant for the purpose of topic modelling in this research are removed.
- Punctuations and numbers are removed as they do not provide any meaningful information about the topics of the research papers being analysed.
- Tokenisation, which is the process of breaking down the string data into individual words, is performed on the data.
- It is noted that these steps are very similar to the ones performed for several research papers discussed earlier in Section 2.

D. Data Mining:

- Latent Dirichlet Allocation (LDA) using Gibbs sampling is applied on the transformed dataset for topic modelling and extracting topics from paper abstracts.
- The parameters passed are 200 omitted Gibbs iterations at the start, 500 Gibbs iterations and 5 repeated random starts.
- Parameters of the symmetric Dirichlet priors were set as alpha=0.1 and beta=0.05. The former results in documents mapped to fewer topics while the latter results in relatively more separated topics.
- Multiple LDA models are trained with varying values for k, the number of topics to be extracted.

E. Interpretation/Evaluation:

- The topic coherence, log likelihoods and R squared values are used as performance metrics to compare the performance of the different LDA models obtained.
- The model with the highest log likelihood and topic coherence is chosen to be the best performer. The evaluation and interpretation is further discussed in more detail in Section 5.

4 Design Specification & Implementation

This section provides an overview of the algorithm used for the purpose of this research.

Latent Dirichlet Allocation or LDA is a generative probabilistic model for mixtures of discrete data like a text corpus (Blei, et al., 2003). It is a three tier hierarchical Bayesian model that considers every document or data item as a large finite mixture over a number of topics. Similarly, every topic is modelled as an infinite mixture over the topic probabilities. LDA differs from simple Dirichlet multinomial-clustering model as the simple model would only be a two tier model where a Dirichlet is sampled once for the textual dataset, a multinomial clustering variable is selected once for each document in the dataset and several words are chosen for the document based on the cluster variable. This is quite similar to several other clustering models because this too involves each document being mapped to only a single topic as opposed to multiple topics as in LDA. LDA is also a three-level model as opposed to two-level and the topic nodes are sampled repeatedly instead of just once.

Gibbs sampling is one of the algorithms in the Markov Chain Monte Carlo (MCMC) framework (Gilks, et al., 1995). It is based on the sampling from conditional distributions of variables. The MCMC algorithms seek to create a Markov chain which has the target posterior distribution as its stationary distribution. This basically means that after traversing through the chain for some iterations, sampling from the distribution must converge to be near sampling from the desired posterior.

5 Evaluation

The LDA models are run for various values for the number of topics. Particularly, models for 5, 10, 20, 30, 40, 50, 100, 150 and 200 topics were run. For evaluating the performance of these models and to select the best model, three metrics are used. These are:

- R^2
- Coherence
- Log Likelihood

It is noted that the best model will be the one with the largest value for R^2, coherence and magnitude of the log-likelihood.

```
> r2_vals
       k=5       k=10       k=20       k=30       k=40       k=50      k=100      k=150      k=200
0.02945186 0.04203987 0.06093408 0.07867645 0.09086205 0.10144587 0.14408750 0.16752303 0.18689378
```

Figure 4: R^2 values for models with different k

```
> summary(model5$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03478 0.04122 0.04881 0.06749 0.08092 0.13174
> summary(model10$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02335 0.02729 0.04915 0.06473 0.08549 0.17608
> summary(model20$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03563 0.05134 0.07482 0.09806 0.13755 0.20699
> summary(model30$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03326 0.05589 0.08084 0.11362 0.14839 0.32405
> summary(model40$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03125 0.06232 0.10893 0.11353 0.14092 0.28979
> summary(model50$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01859 0.05711 0.09997 0.11806 0.14102 0.46034
> summary(model100$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01875 0.06145 0.09833 0.11967 0.15908 0.46034
> summary(model150$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01122 0.06639 0.10672 0.11888 0.14962 0.46034
> summary(model200$coherence)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00963 0.05999 0.09277 0.11051 0.14364 0.46034
```

Figure 5: Coherence for model with different k

```
> model5$log_likelihood[50,]
   iteration log_likelihood
50       490      -16349007
> model10$log_likelihood[50,]
   iteration log_likelihood
50       490      -16913107
> model20$log_likelihood[50,]
   iteration log_likelihood
50       490      -17508800
> model30$log_likelihood[50,]
   iteration log_likelihood
50       490      -17936538
> model40$log_likelihood[50,]
   iteration log_likelihood
50       490      -18218150
> model50$log_likelihood[50,]
   iteration log_likelihood
50       490      -18483745
> model100$log_likelihood[50,]
   iteration log_likelihood
50       490      -19243241
> model150$log_likelihood[50,]
   iteration log_likelihood
50       490      -19764247
> model200$log_likelihood[50,]
   iteration log_likelihood
50       490      -20225975
```

Figure 6: Log-likelihood for models with different k

It is evident from figures 4 and 6 that the values for R^2 and log-likelihood are increasing with increasing k thereby portraying direct proportionality in this case. Coherence also shows a similar trend but with some exceptions.

We discard the values of k greater than 50 because the greater the number of topics, the lesser will be the clarity/understanding and verifiability by experts (Battisti, et al., 2015). Upon discarding values greater than 50, it is observed that k=50 is the optimal number of topics for our dataset containing 14,116 abstracts.

It is also noted that topic prevalence, which is simply the prevalence of the topic, should be directly proportional to alpha. This is because if a topic is very prevalent, then there will be few other topics apart from that. Fewer topics are associated with higher values for alpha. This proportionality can be confirmed by figure 7 as seen below:

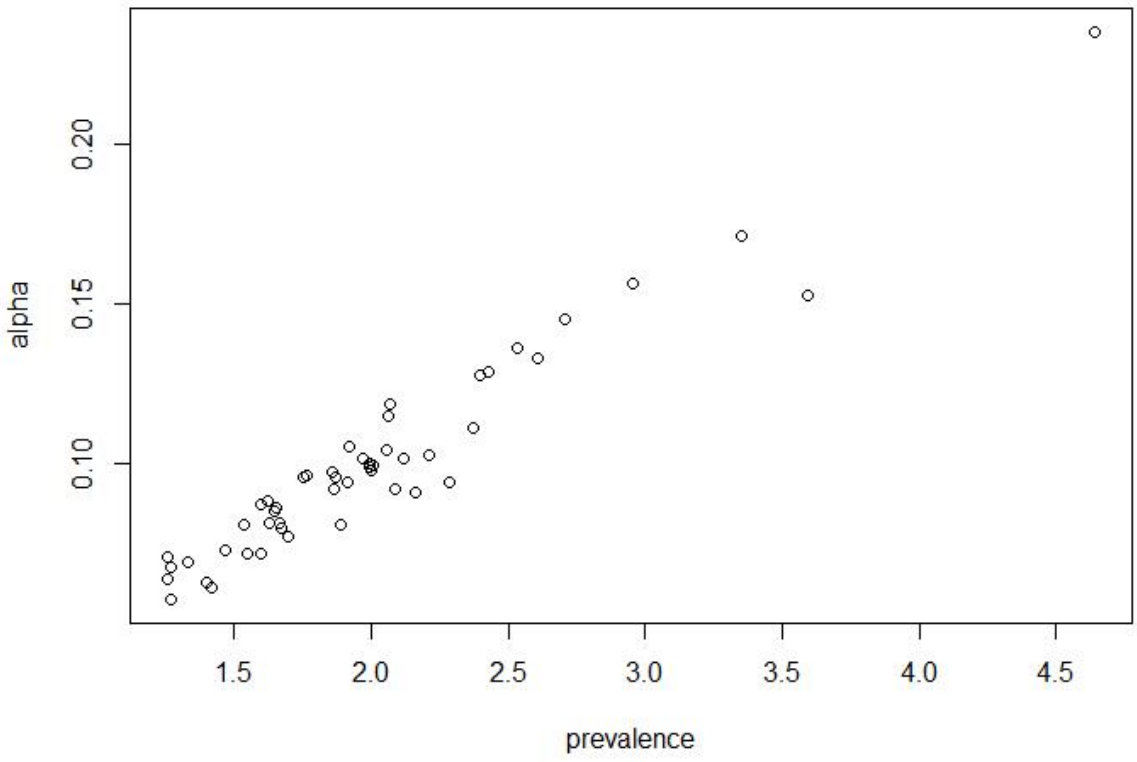


Figure 7: Prevalence vs Alpha

The R library textmineR is used to create labels for the topics identified by the LDA model. The fifty labelled topics identified can be seen below in figure 8:

```
t_1  "software_development"
t_2  "information_technology"
t_3  "case_study"
t_4  "developing_countries"
t_5  "oil_gas"
t_6  "financial_services"
t_7  "supply_chain"
t_8  "proposed_method"
t_9  "proposed_scheme"
t_10 "game_theory"
t_11 "higher_education"
t_12 "power_consumption"
t_13 "regulatory_framework"
t_14 "management_system"
t_15 "time_series"
t_16 "smart_contracts"
t_17 "renewable_energy"
t_18 "science_technology"
t_19 "monte_carlo"
t_20 "social_network"
t_21 "cloud_computing"
t_22 "wi_fi"
t_23 "neural_network"
t_24 "case_study"
t_25 "health_care"
t_26 "financial_crisis"
t_27 "regulatory_networks"
t_28 "electric_vehicles"
t_29 "decision_making"
t_30 "financial_performance"
t_31 "mobile_payment"
t_32 "blockchain_technology"
t_33 "neural_network"
t_34 "real_time"
t_35 "power_system"
t_36 "internet_things"
t_37 "social_media"
t_38 "smart_grid"
t_39 "big_data"
t_40 "cyber_attacks"
t_41 "blockchain_technology"
t_42 "listed_companies"
t_43 "blockchain_technology"
t_44 "real_time"
t_45 "machine_learning"
t_46 "data_mining"
t_47 "access_control"
t_48 "real_estate"
t_49 "stock_market"
t_50 "gene_regulatory"
```

Figure 8: Labelled Topics

Top 10 topics along with their labels, coherence, prevalence and the top 10 words in each topic are shown below in figure 9:

```
> model50$summary[ order(model50$summary$prevalence, decreasing = TRUE) , ][ 1:10 , ]
    topic              label_1 coherence prevalence
t_41  t_41 blockchain_technology     0.025      4.643
t_26  t_26      financial_crisis     0.042      3.596
t_43  t_43 blockchain_technology     0.143      3.350
t_38  t_38            smart_grid     0.188      2.954
t_24  t_24             case_study     0.040      2.709
t_30  t_30 financial_performance     0.044      2.609
t_17  t_17       renewable_energy     0.176      2.536
t_47  t_47         access_control     0.102      2.426
t_8    t_8       proposed_method     0.113      2.397
t_29  t_29       decision_making     0.108      2.375
                                                                                                                top terms
t_41                  technology, research, technologies, paper, challenges, future, recent, years, applications, industry
t_26                    financial, development, china, paper, industry, enterprises, economic, crisis, financing, economy
t_43 blockchain, technology, distributed, blockchain_technology, system, based, decentralized, trust, data, blockchain_based
t_38                                  energy, grid, demand, electricity, market, power, smart, load, smart_grid, electric
t_24                                model, systems, approach, based, paper, architecture, framework, system, models, agent
t_30                           study, research, factors, performance, results, analysis, data, model, findings, influence
t_17                         energy, power, renewable, wind, system, solar, generation, renewable_energy, pv, electricity
t_47                              security, privacy, data, access, authentication, secure, user, information, scheme, key
t_8                          problem, algorithm, proposed, optimization, method, optimal, based, results, approach, model
t_29                       risk, evaluation, decision, financial, making, model, method, performance, decision_making, risks
```

Figure 9: Top 10 topics

The following pie-chart shows the top 15 topics and their respective prevalence to give an overview of which topics were the most discussed in academic literature within the FinTech domain:

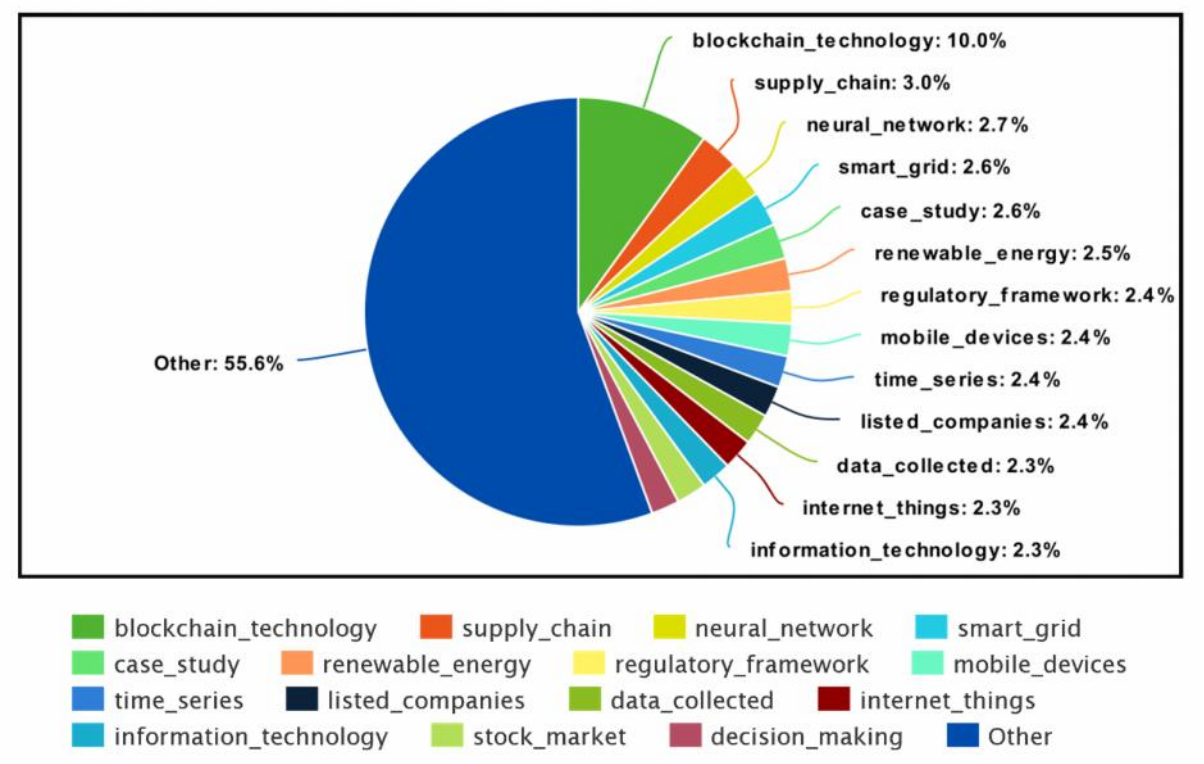


Figure 10: Top 15 Topics and their prevalence

The following figure 11 shows the frequency of documents in each year since 2010 so as to provide an idea about how many publications existed for each year.

```
> doc_freq

          2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Frequency 1658 1632 1060  968 1030 1138 1198 1593 2724 1115
```

Figure 11: Document frequency by year

It is observed that there are no anomalies and the number of publications do not vary substantially over the years.

The topics are converted to time series and their slope is calculated to determine whether the topics have positive (increasing) or negative (decreasing) slope. The figure 12 below shows a table of number of topics with positive and negative trends by their level of significance.

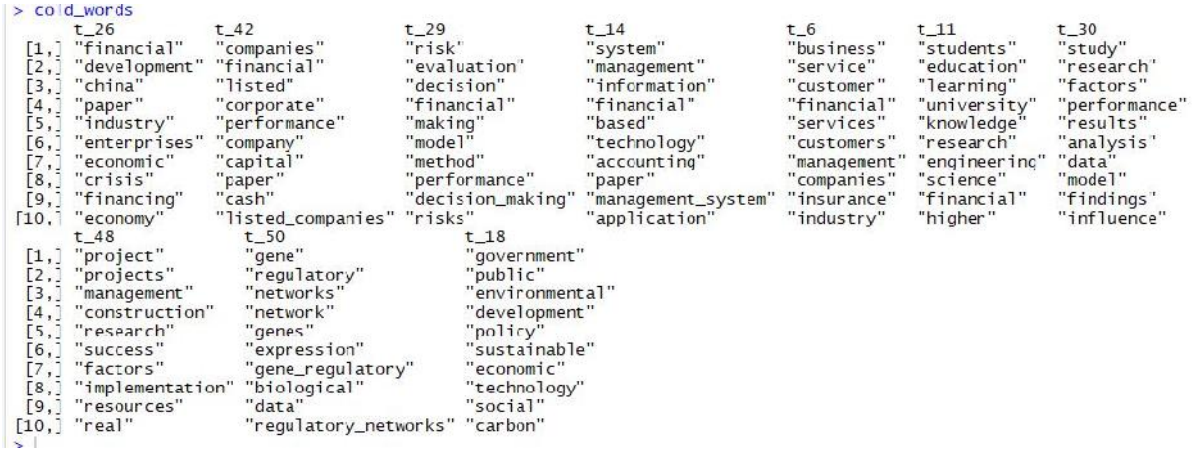
	p<0.05	p<0.01	p<0.001	p<0.0001
Negative trend	13	11	6	4
Positive trend	9	8	2	0
Total	22	19	8	4

Figure 12: Linear trends by significance

The top hot topics, that is, the topics which are increasing in popularity as well as the cold topics (decreasing in popularity) are also determined. These hot and cold topics, along with the top words in each topic are displayed below in figures 13 and 14 respectively.

```
> hot_words
      t_43                  t_32           t_36             t_47           t_16              t_40          t_41
 [1,] "blockchain"          "bitcoin"      "iot"            "security"     "smart"           "security"    "technology"
 [2,] "technology"          "blockchain"   "devices"        "privacy"      "contracts"       "attacks"     "research"
 [3,] "distributed"         "transactions" "internet"       "data"         "contract"        "cyber"       "technologies"
 [4,] "blockchain_technology" "transaction" "things"        "access"       "smart_contracts" "attack"      "paper"
 [5,] "system"              "consensus"    "internet_things" "authentication" "smart_contract" "detection"  "challenges"
 [6,] "based"               "peer"         "security"       "secure"       "ethereum"        "financial"   "future"
 [7,] "decentralized"       "block"        "smart"          "user"         "blockchain"      "malicious"   "recent"
 [8,] "trust"               "cryptocurrency" "things_iot"   "information"  "platform"        "threats"     "years"
 [9,] "data"                "mining"       "iot_devices"    "scheme"       "based"           "malware"     "applications"
[10,] "blockchain_based"    "currency"     "edge"           "key"          "execution"       "detect"      "industry"
      t_39          t_9
 [1,] "data"        "proposed"
 [2,] "big"         "scheme"
 [3,] "big_data"    "based"
 [4,] "analysis"    "signature"
 [5,] "large"       "dit"
 [6,] "processing"  "paper"
 [7,] "analytics"   "schemes"
 [8,] "storage"     "compared"
 [9,] "amount"      "performance"
[10,] "information" "codes"
```

Figure 13: Hot Terms/words



```
> cold_words
      t_26           t_42               t_29             t_14                 t_6           t_11          t_30
 [1,] "financial"    "companies"        "risk"           "system"             "business"    "students"    "study"
 [2,] "development"  "financial"        "evaluation"     "management"         "service"     "education"   "research"
 [3,] "china"        "listed"           "decision"       "information"        "customer"    "learning"    "factors"
 [4,] "paper"        "corporate"        "financial"      "financial"          "financial"   "university"  "performance"
 [5,] "industry"     "performance"      "making"         "based"              "services"    "knowledge"   "results"
 [6,] "enterprises"  "company"          "model"          "technology"         "customers"   "research"    "analysis"
 [7,] "economic"     "capital"          "method"         "accounting"         "management"  "engineering" "data"
 [8,] "crisis"       "paper"            "performance"    "paper"              "companies"   "science"     "model"
 [9,] "financing"    "cash"             "decision_making" "management_system"  "insurance"   "financial"   "findings"
[10,] "economy"      "listed_companies" "risks"          "application"        "industry"    "higher"      "influence"
      t_48            t_50              t_18
 [1,] "project"       "gene"            "government"
 [2,] "projects"      "regulatory"      "public"
 [3,] "management"    "networks"        "environmental"
 [4,] "construction"  "network"         "development"
 [5,] "research"      "genes"           "policy"
 [6,] "success"       "expression"      "sustainable"
 [7,] "factors"       "gene_regulatory" "economic"
 [8,] "implementation" "biological"     "technology"
 [9,] "resources"     "data"            "social"
[10,] "real"          "regulatory_networks" "carbon"
>
```

Figure 14: Cold Terms/words

It is observed that blockchain technology, internet of things, access control, smart contracts, cyber-attacks, scheme and big data are the hot topics. Moreover, financial crisis, listed companies, decision making, management system, financial services, higher education, financial performance, real estate and science and technology are the cold topics.

So far only linear trends in topics have been considered. This only allows for topics having either a positive or a negative slope. However, in reality, the topics do not vary so uniformly with the passage of time. Due to this reason, there is need to incorporate topics that do not have a linear positive or negative trend. This research overcomes this drawback by also considering non-linear trends in research topics. Multilayer perceptrons (MLPs) with two hidden layers are used to calculate the mean of topic probabilities over all documents for every topic (Bittermann & Fischer, 2018). This average topic probability is then modelled as a non-linear function of the year of publication. The two hidden units mitigate the risk of overfitting and also allow for non-monotonic functions. The topics having non-linear trends are identified and these are displayed below in figure 15:

```
> terms_nonlinear
      t_35           t_17                t_8             t_21              t_27          t_22          t_33
 [1,] "power"        "energy"            "problem"       "cloud"           "control"     "spectrum"    "learning"
 [2,] "system"       "power"             "algorithm"     "service"         "system"      "wireless"    "neural"
 [3,] "distribution" "renewable"         "proposed"      "services"        "controller"  "access"      "network"
 [4,] "voltage"      "wind"              "optimization"  "computing"       "model"       "radio"       "prediction"
 [5,] "transmission" "system"            "method"        "cloud_computing" "time"        "regulatory"  "model"
 [6,] "paper"        "solar"             "optimal"       "providers"       "paper"       "interference" "machine"
 [7,] "grid"         "generation"        "based"         "resources"       "stability"   "channel"     "neural_network"
 [8,] "network"      "renewable_energy"  "results"       "data"            "proposed"    "networks"    "artificial"
 [9,] "load"         "pv"                "approach"      "provider"        "regulatory"  "mobile"      "accuracy"
[10,] "losses"       "electricity"       "model"         "resource"        "simulation"  "paper"       "machine_learning"
      t_24          t_25          t_38
 [1,] "model"       "health"      "energy"
 [2,] "systems"     "medical"     "grid"
 [3,] "approach"    "healthcare"  "demand"
 [4,] "based"       "care"        "electricity"
 [5,] "paper"       "patients"    "market"
 [6,] "architecture" "patient"    "power"
 [7,] "framework"   "insurance"   "smart"
 [8,] "system"      "health_care" "load"
 [9,] "models"      "system"      "smart_grid"
[10,] "agent"       "hospital"    "electric"
```

Figure 15: Non-linear terms

It is observed that power systems, renewable energy, cloud computing, regulatory networks, wifi, neural network, case study, healthcare and smart grid have non-linear trends. These non-linear trends can be visualised as time series. Time series visualisations for topics having non-linear trends are shown below in figure 16:

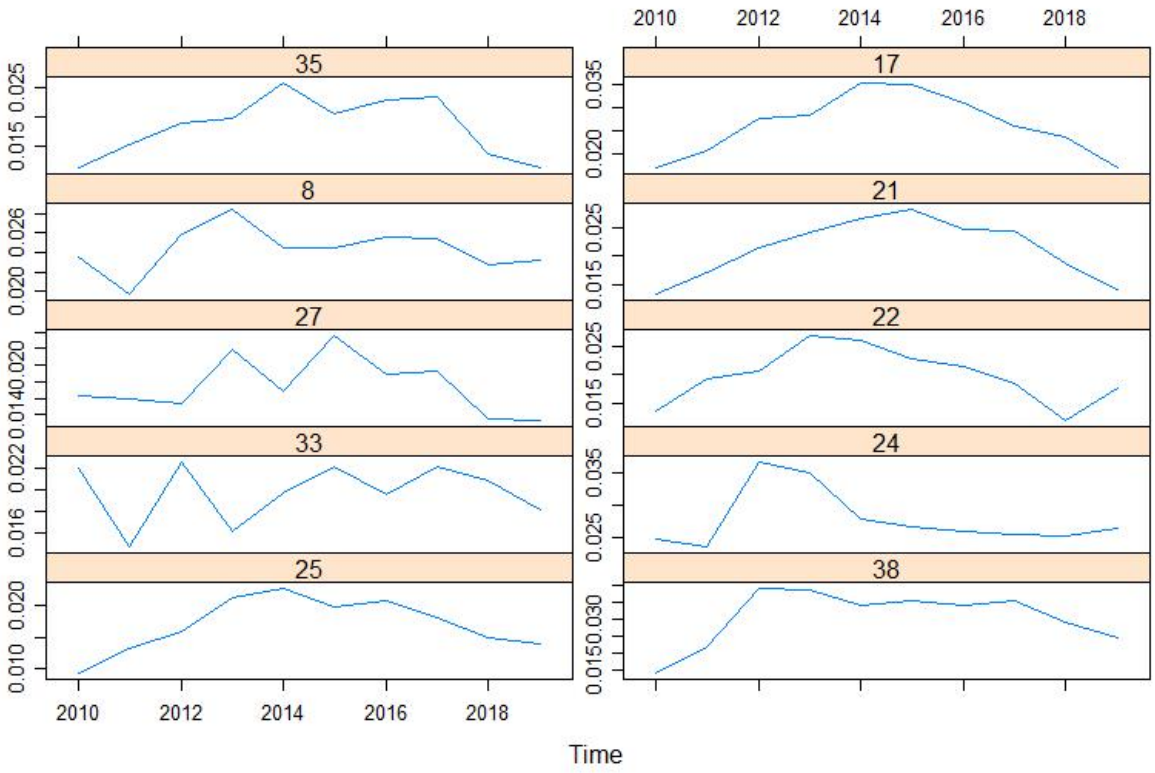


Figure 16: Non-linear topics as time series

6 Conclusion and Future Work

This research analyses research done in the field of financial technology or FinTech since 2010 to identify trends in the topics of research. Latent Dirichlet Allocation using Gibbs sampling is used for topic modelling. 14,116 abstracts from journal articles and conference proceedings are extracted from IEEE Xplore Digital Library and 50 topics are extracted from

this corpus. The number of topics is selected based on values for coherence, log-likelihood and R^2. The values for these metrics increases with higher values of k but since large number of topics reduce understanding and hampers verifiability by experts (Battisti, et al., 2015), values of k greater than 50 are discarded. It is observed that among the 50 topics identified, some topics portray increasing linear trends over time including blockchain technology, internet of things, access control, smart contracts, cyber-attacks, scheme and big data. These topics are recognised as hot topics in the field of FinTech. The hot topics identified in this research will be more useful to new entrepreneurs and businesses who can use this knowledge about existing trends and popular technologies to facilitate their use cases. It should be noted that from the dataset used for this research, the hot topics identified are the topics that have already been fairly popular in recent years. Therefore, it can be concluded that this type of research needs to be done regularly with more data so as to identify hot topics that are yet to gain popularity. This means that students of FinTech, business leaders in the FinTech domain and governments/policy makers must give special attention to these topics as they are likely to become more widespread in the near future. Business leaders should push for research into these topics by their R&D department and policy makers should learn about these topics in order to be able to regulate the novel technologies. Students should also delve into these to see if they pique their interest. Moreover, some cold topics are also identified which can be avoided by the stakeholders mentioned above since their popularity is dropping. Unlike many topic modelling implementations, this research also identifies topics having non-linear trends using MLPs. These include topics like power systems, renewable energy, cloud computing, regulatory networks, wifi, neural network, case study, healthcare and smart grid. These are visualised as time series for easier grasping. This research provides valuable insight to students, policymakers and business leaders. It can be concluded that topic modelling using LDA is a feasible method for exploratory analysis of the research topics in the FinTech domain as well as determining the trends in these topics.

Although this research provides many benefits over several topic modelling implementations based on LDA, there is still room for much improvement. For future analysis, several other data sources can be included to provide a more holistic representation of the FinTech research done around the globe. Furthermore, more sophisticated techniques like correlated topic models and dynamic topic modelling can be used to improve the results of this research.

References

Abuhay, T. M., Nigatie, Y. G. & Kovalchuk, S. V., 2018. Towards predicting trend of scientific research topics using topic modeling. Heraklion, Greece, 7th International Young Scientist Conference on Computational Science.

Abulaish, M. and Fazil, M. (2018). Modeling Topic Evolution in Twitter: An Embedding-Based Approach. *IEEE Access*, 6, pp.64847-64857.

Amado, A., Cortez, P., Rita, P. and Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, 24(1), pp.1-7.

Battisti, F. D., Ferrara, A. & Salini, S., 2015. A decade of research in statistics: a topic model approach. Scientometrics, 103(2), p. 413–433.

Bittermann, A. and Fischer, A. (2018). How to Identify Hot Topics in Psychology Using Topic Modeling. *Zeitschrift für Psychologie*, 226(1), pp.3-13.

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent dirichlet allocation. The Journal of Machine Learning Research, Volume 3, pp. 993-1022.

Chanda, P. & Das, A. K., 2018. A novel graph based clustering approach to document topic modelling. Bangalore, India, 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

Deerwester, S. et al., 1990. Indexing by latent semantic analysis. Journal of the association for information science and technology, 41(6), pp. 391-407.

Deloitte, 2017. Fintech by the numbers. [Online] Available at: https://www2.deloitte.com/tr/en/pages/financial-services/articles/fintech-by-the-numbers.html#
[Accessed 11 August 2019].

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), pp.27-34.

Feldman, R. & Dagan, I., 1995. Knowledge discovery in textual database (KDT). s.l., Proceedings of the first ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 112-117.

Gilks, W., Richardson, S. & Spiegelhalter, D., 1995. Markov Chain Monte Carlo in Practice. 1 ed. s.l.:Chapman and Hall/CRC.

Hidayatullah, A. F. et al., 2018. Twitter topic modeling on football news. Nagoya, Japan, 2018 3rd International Conference on Computer and Communication Systems (ICCCS).

Hofmann, T., 1999. Probabilistic latent semantic analysis. Stockholm, Sweden, UAI'99 Proceedings of the fifteenth conference on uncertainty in artificial intelligence.

IEEE, 2019. IEEE Xplore. [Online] Available at: https://ieeexplore.ieee.org/Xplore/home.jsp
[Accessed 11 August 2019].

Johri, V. & Bansal, S., 2018. Identifying trends in technologies and programming languages using topic modeing. Laguna Hills, CA, USA, 2018 IEEE 12th International Conference on Semantic Computing (ICSC).

KPMG, 2019. Total value of investments into Fintech companies worldwide from 2010 to 2018 (in billion U.S. dollars). [Online] Available at: https://www.statista.com/statistics/719385/investments-into-fintech-companies-globally/
[Accessed 11 August 2019].

Laoh, E., Surjandari, I. & Febirautami, L. R., 2018. Indonesians' song lyrics topic modelling using latent dirichlet allocation. Zhengzhou, China, 2018 5th International Conference on Information Science and Control Engineering (ICISCE).

Lee, J., Kang, J., Jun, S., Lim, H., Jang, D. and Park, S. (2018). Ensemble Modeling for Sustainable Technology Transfer. *Sustainability*, 10(7), p.2278.

Liu, Z. et al., 2018. An emotion oriented topic modeling approach to discover what students are concerned about in course forums. Mumbai, India, 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT).

McAuley, J. & Leskovec, J., 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. Hong Kong, China, RecSys '13 Proceedings of the 7th ACM conference on Recommender systems.

Vamshi, K. B., Pandey, A. K. & Siva, K. A. P., 2018. Topic model based opinion mining and sentiment analysis. Coimbatore, India, 2018 International Conference on Computer Communication and Informatics (ICCCI).

Wai, T. and Aung, S. (2018). TPMTM: Topic Modeling over Papers' Abstract. *Advances in Science, Technology and Engineering Systems Journal*, 3(2), pp.69-73.

Wang, J. C., Pan, J. G. & Zhang, F. Y., 2000. Research on web text mining. Journal Of Computer Research And Development.

Wang, X. & Yang, B., 2018. STMF: A sentiment topic matrix factorization model for recommendation. Nagoya, Japan, 2018 3rd International Conference on Computer and Communication Systems (ICCCS).