[[1]](#footnote-1)

An Evaluation of Topic Models for the Estimation of Unobserved Variables in Structured and Unstructured Documents

Collins N. Udanor, Onyinye E. Nweke, *Graduate Student Member, IEEE,* and George E. Okereke

***Abstract*—For effective data collection, researchers are often faced with three challenges of where, what and how? Where to find researchable data, with what tools and methodologies to scrape websites for such data, and how to perform the required analytics and extract insightful knowledge. This study examines the possibility and the extent user tweets could influence the direction of research, especially in the field of machine learning and artificial intelligence. In this paper, we use the Latent Dirichlet Allocation (LDA) topic modelling technique to discover machine learning research topics popularity in 35,860 unorganized datasets (tweets) from 20 Artificial Intelligence and machine learning related handles, while using 7,241 articles from 42 years’ Neural Information Processing Systems (NIPS) conference papers dataset, an organized document as a control. The Latent Semantic Index (LSI) and the Hierarchical Dirichlet Process (HDP) are used to compare the performance of the LDA. Embedding methods such as the bag of words and the term frequency inverse document frequency (tf-idf) are used to encode the corpora and compared. Results suggest that using the structured dataset guaranteed better classification, though the unstructured dataset is quite informative. However, a t-test showed that the difference between the results of the two datasets was not significant. The LDA model consistently out-performed LSI and HDP across topics, respectively. A comparison of the Gensim and Mallet Python frameworks showed that Mallet promised a better topic modelling result than Gensim.**

***Index Terms*— Bag of words, Gensim, Latent Dirichlet Allocation, Machine learning, Mallet, Topic models, Twitter.**

# I. INTRODUCTION

T

HERE is no doubt that the social media has impacted on every sphere of our daily, recreational, economic, political, and research lives. It has also become a medium for the generation and curation of research data, as we see tonnes of data explosion in the cyber space due to people’s interactions with one another through the social media, or with machines in e-commerce websites, through online gaming, sports, and the like. Researchers are leveraging on these new data highways to study human behaviour, buying patterns, research trends, opinions pools, and other outcomes, as well as monitor early warning signs from an individual’s health, earthquakes and other natural disasters.

Twitter’s rapid content delivery, dissemination, and the provision of Search and Stream Application Programmers Interfaces (APIs) seems to have placed Twitter in the number one spot for research data collection, above other social media sites, like Facebook, and the rest. Unlike Facebook, less than 10% of Twitter’s accounts are private, making it easier for researchers to collect data from the over 90% public Twitter [1]. Twitter is seen as the “SMS” of the social media, particularly because of the restriction on number of characters and its rapid content delivery. The popularity of Twitter in the academia has led to the proposal on developing alternative metric (altmetric), the t-factor for measuring a researcher’s impact based on retweets [2-4].

The effect of Twitter on academic journals and research papers was analysed by Bornmann and Haunschild [2], considering paper dissemination via tweets and their impacts through citations. The outcome of their study shows a greater amount of citations of journals with twitter account as compared to those without social media (Twitter’s) presence. The researchers sought to find out the best method for journals to participate on twitter. Twitter account metrics such as total tweets, followers and followings were considered to find out the rate at which they could influence citations. In addition, an alternative metrics provider called Plum Analytics was also used. Two statistical analysis i.e. student’s T-test and regression analysis were applied on a dataset of 4,176 research articles found in 350 journals. The journals were classified according to four possibilities (journal account, owner account, publisher account, no twitter account). The study concludes that the number of followers on a Twitter profile majorly influences the number of tweets and citations a particular paper receives.

A t-factor for measuring the impacts of publications on Twitter based on the popular h-index, which is a metric used for measuring research performance was proposed by Ortega [3]. The authors made use of a single paper from Scopus. They made use of the formula for h-index as a technique to combine all tweets and retweets of a particular researcher or paper publication into a single tweet for the purpose of analysis.

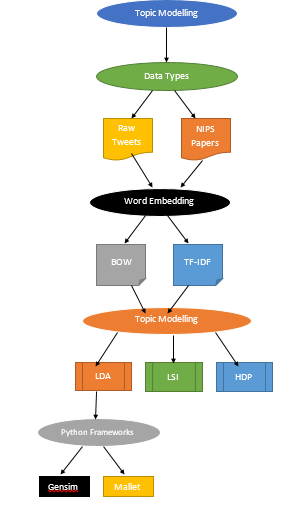
An overview of Twitter activity as the basis of scholarly twitter metrics was given by Haustein [4]. The author analysed 24 million tweets with the intention of analysing tweets measurable with respect to research evaluation. Their study finds that majority of tweets related to research output are mostly created by the academia, hence the impact is more on academic communication rather than societal. They also observed that the peak of transmission of these academic tweets happen shortly after publication. There is however a low discussion of academic tweets generally among twitter users. Tweets, if properly analysed can reveal trending topics of discussions among Twitter users, and provide insights to the general perception of the groups involved in the tweet, retweet, and mentions.

A topic is semantically defined as co-occurrence patterns of words in the context of tweets, sentence, paragraph, news articles or book chapters. A document is decomposed into topics and words, and modelled. Several models for topic modelling abound including; the Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), Spherical HDP, Latent Semantic Analysis/Latent Semantic Indexing (LSI), pLSA, and so on. The Latent Dirichlet Allocation (LDA) considers a document as a collection of topics in certain proportions. It is a bag of words model which was developed in 2003 and has several fields of applications such as document classification, sentiment analysis, bioinformatics, etc. The observable features seen by the model are the words that appear in documents.

This work uses unstructured and structured datasets, and topic modelling techniques to discover the trending machine learning and AI topics, and how they influence machine learning research.

In this work we demonstrate document modelling to discover hidden topics in raw tweets from Twitter microblog and machine learning papers from the NIPS conferences between 1975 and 2017 from the Kaggle dataset repositories. The datasets are modelled using three topic modelling techniques; LDA, LSI, and HDP. The performances are compared to determine the best algorithm. Though the LDA model which is our benchmark for topic modelling uses the bag of words (BOW) embedding, this research attempts to find if there is any significant difference in result if the Term Frequency Inverse Document Frequency (TF-IDF) embedding is used instead of the bag of words (BOW).

Fig. 1 gives a graphical overview of the research processes in this work. Section 2 gives a brief background into topic modelling techniques and how they work. The section also reviews related works done with Twitter data, with multiple topic modelling techniques and concludes with a review of topic modelling techniques. Section 3 lays out the experimental works, showing the two major types of data, tweets and NIPS papers used for the experiment. The data pre-processing and transformation methods used are also described. The experiment with LDA, LSI, and HDP topic modelling are also described, as well as the data embedding techniques, and the Python frameworks used. Section 4 presents four different results from each of the experimental stages, while Section 5 discusses the findings and compares them with the ground truths.



**Fig. 1.** The workflow of this research.

# II. Background

Text mining is a process of discovering hidden knowledge from a corpus of text which may contain different topics using one or more text mining or machine learning algorithms. The output of the text mining task is a term document matrix (TDM). The TDM serves as an input into the topic model, which is a generative process in which a machine learns automatically from unlabelled documents in an unsupervised way. The author in [5] was among the first to propose topic model with the probabilistic Latent Semantic Analysis (pLSA). Among the most popular document modelling is the Latent Dirichlet Allocation (LDA) [6], a Bayesian mixed method for analysing discrete data for topics that were thought to be uncorrelated. The correlated topic model (CTM) is an extension of the LDA, in which correlations between topics are allowed. In machine learning and natural language processing, topic models provide a framework for the term frequency occurrence in documents in a corpus [7]. Topic models are a mixed membership models so they differ from unigrams or the mixture of unigram models.

1. *How LDA Works*

The LDA considers a document as a collection of topics in certain proportions. It is a bag of words model which was developed in 2003 and has several fields of applications such as document classification, sentiment analysis, bioinformatics, etc. The observable features seen by the model are the words that appear in documents. Other parameters are latent (hidden), one of such is a topic that is assigned to those words, thereby making every document a mixture of such topics. Fig. 2 shows how the model combines the three entities; document, topics and words. M is the number of documents, N is the number of topics, and θ is the distribution of each of such topics. A single document has N words, and each word has N topics.

M

N

**Fig. 2.** Plate notation representation of the LDA model

Each document can be described by what topics it contains and by what percentage they contain such topics. For example, a document may contain 50% of AI, 30% of OS, and 10% gaming, etc. The model will take such document and generate a new document by taking the right number of words from specific topics and put them together.

LDA as a bag of words model in which every document is a collection of words taken from different buckets or topics depending on what it writes about. We need to specify the number of topics to generate from a document and give rules on how they should be constructed. These rules are hyper-parameters, represented by alpha and beta. These are both parameters of the Dirichlet distribution.

Alpha controls per document topic distribution, and beta is responsible for per topic word distribution. When a high alpha value is obtained it implies that a document is likely to contain a mixture of most of the topics, not just a single topic. Conversely, a low alpha value means that a document is most likely to contain just a few topics. On the other hand, a high beta value indicates that each topic is likely to contain a mixture of most of the words, while a low beta value means a topic will most likely contain a mixture of a few words. In summary, a high alpha value will make documents appear more similar to each other and a high beta value will make topics appear similar to each other. The following formula captures all what we have said so far. This is the total probability of the model, given by equation (1);

P(W, Z, q, j, a, b) = ) Jj;a)Zj,t;J)P(Wj,t|j) (1)

There is no specific number of topics to look for, it all depends on the application and use case.

1. **Model Specification for LDA**

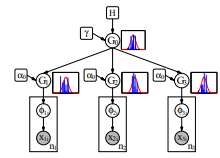
As stated earlier, LDA uses generative processes involving three steps for topic modelling.

1. Term document distribution determined by: ~ Dirichlet(
2. The proportion of topics q of the topic distribution for the document W, determined by: q ~ Dirichlet (a).
3. For each of N words Wi the topics, we can;
4. Choose a topic Zi ~ multinomial(q)
5. Choose a word Wi from a multinomial probabilistic distribution which is conditioned on the topic,

Zi: P(Wi|Zi,b), where b is the term distribution of topics and also contains the probability of word occurring in a given topic.

1. **Hierarchical Dirichlet Process (HDP**)

HDPis based on the dirichlet process mixture model where the dirichlet process for each group shares a base distribution. HDP is used for clustering grouped data which are usually non parametric. HDP can be applied as a topic model, where words are organized into documents and each document is made up of a bag of words. The number of states in HDP is not bounded and can be learnt from the data, it is a major component of the infinite hidden markov model. Fig. 3 illustrates the HDP process.



**Fig. 3.** Graphical model of HDP with three groups

1. **Latent Semantic Indexing (LSI) -** LSI model uses a mathematical technique called singular value decomposition (SVD). Along with reducing the number of rows, it also preserves the similarity structure among columns. In matrix, the rows represent unique words and the columns represent each document. It works based on distributional hypothesis, i.e. it assumes that the words that are close in meaning will occur in same kind of text. LSI uses statistically derived conceptual indices for retrieval rather than individual words, thereby overcoming the challenges that comes with lexical matching. LSI is based on the principles that words that are used together or in the same context have similar meaning. In other words it assumes that synonymous words are often used together.

*B. Review of Related Literature*

This section reviews related works done with Twitter data, with multiple topic modelling techniques and concludes with a review of topic modelling techniques.

1. **Twitter Data Researches**

Microblogging websites such as Twitter has been the centre of attraction for many data mining researches, especially in the area of sentiment analysis and opinion mining. Different techniques and tools have been applied to analyse these large data sets to find meaningful trends and make predictions and forecast in areas that interest the researchers. These techniques include but not limited to machine learning and statistical methods. Different data mining tools are being explored by researchers in order to effectively manage these data. Several topics have been covered by researchers using Twitter data as source of input for the research. Some areas where these researches have focused in previous years are sentiment analysis for disease detection, prediction [8-10] and monitoring [11], [12]. Also for election prediction in [13-18]. Another notable area is that of stock market prediction as seen in the works of [19-21]. Other notable areas are crime prediction [22] and spam detection [23].

1. **Disease Detection:** In order to evaluate the use of twitter data for predicting the level of flu cases in New York City, researchers in [8] applied techniques such as linear regression, Poisson model and vector map to a collection of twitter dataset gathered from users in the proposed area. For the purpose of monitoring flu cases, [11] used geographic information system (GIS) to target, filter and normalize data. Also, they applied a comprehensive data mining process using support vector machine (SVM) as a classifier. Meanwhile, the authors in [12] extended the research on twitter data from prediction and monitoring of flu cases to monitoring cancer patients and the treatments they use. The researchers in [9], were able to forecast the onset of mental illness such as depression and post-traumatic stress disorder in twitter users using Random forest classifier, with Python and R for the analysis. Agglomerative clustering was proposed by [10] for early detection of disease outbreak, this was to help manage such events effectively.
2. **Election Prediction:** Statistical tools with relative support parameters where used by [13] to analyse tweets from Italian users. They compared tweets by party leaders to the final election result. They found out that the volume and rate of tweets have huge effect on the election result.

To predict the outcome of both the US Presidential Elections of 2012 and Karnataka Assembly Elections 2013, [14] combined principle component analysis (PCA) and support vector machine (SVM) for the analyses. Other techniques employed in this study include Naive Bayes, Maximum Entropy and Artificial Neural Networks. Their experimental results showed that Support Vector Machine supersedes other techniques investigated in the prediction of results. Contradicting previous report, [15] found no correlation between the analysis result of twitter data and the election outcomes while analysing the twitter dataset related to the 2010 US congressional elections. In [16], a real time sentiment analysis of public view on the US 2012 presidential election using Naive Bayes for tweet classification was proposed. The research conducted in [17], [18] used sentiment analysis to predict the 2017 French presidential election and the 2011 Singapore presidential elections, respectively. The former extracted data from microblog twitter to predict the popularity of candidates, while the latter uses weighting techniques to predict the percentage of votes each candidates will receive.

1. **Stock Market Prediction:** A system was designed and implemented by Skuza and Romanowski [19], for the purpose of stock market prediction based on the analysis of twitter data. Naive Bayes as a machine learning technique was applied for classification of tweets collected over a period of three months. Also, [20] applied linear regression with an exogenous data (twitter dataset) model to predict stock market indicators. Their preliminary results showed that the daily number of tweets had a correlation with certain stock market indicators at each level. The authors in [21] applied sentiment analysis and supervised machine learning principles to the tweets extracted from twitter and analysed the correlation between stock market movements of a company and sentiments in tweets. There results showed that there is a correlation between the price of stock market and the tweets of users.
2. **Topic Modelling Researches**

Topic modelling finds its application in many diverse fields of life endeavour. One of such is in news, and poetry like the Urdu text modelling [24] in which the LDA, LSI and HDP models were applied and compared. Findings showed that LDA did better than LSI, and both did well with the news corpus text, but none of the models excelled in the raw poetry data. Another area of topic modelling application is in legal document, such as in the case of the Latvian legal document [25]. LDA, LSI, and HDP were also employed in the modelling and compared. Though, they all yielded negative coherence values, which was used as a measure of goodness, the LDA slightly out-performed the other two. Legal document modelling in the British legislation [26], was also modelled using LDA, HDP, and Saffron, a domain modelling software. Saffron excelled over LDA, and HDP in that order. Topic modelling can also be used to extract topics from a multilingual survey corpus, and automatically classify them [27]. LDA and HDP were used, and HDP was able to automatically group a good number of topics. Topic modelling using LDA is also useful in image and music domain [28].

1. *Review of Document Modeling Techniques*

Microblogging data like that from Twitter requires a hierarchical model such as the LDA that share structure between documents where objects of interest exhibit structural representation. The frequency of topics in corpora follow the power-law distribution where a few topics which occur more frequently have higher probability with the majority of the topics having low probability. This leaves an imbalance. A modified version of the Restricted Boltzmann Machine (RBM), the Diversified RBM (DRBM) promises to address this problem [29] by diversifying the hidden units to represent not only the dominant topics but also the less dominant ones in the long tail region. The LDA, however, uses a probabilistic Dirichlet-multinomial setting which is unable to capture the power-law distribution, also known as the Zipf’s laws in linguistics, produced by a stochastic process with frequent outcomes that attract probability mass such as the “rich-get-richer” process [30]. The authors therefore proposed a novel stochastic topic modelling approach based on the hierarchical Pitman-Yor (PY) process for word distribution modelling. LDA and PY differ in the way topics are generated in a document. Whereas, LDA generates a topic from each word in a document, in PY a topic is generated in a table in what is called the Chinese Restaurant Process (CRP) in the PY model. A variation of the CRP is the Nested Chinese Restaurant Process (nCRP), a derivative of the Hierarchical Dirichlet Process (HDP). Ahmed, Hong, & Smola [31] proposed a Chinese Restaurant Franchise (CRF) model which they used to model the joint distribution of the Twitter microblog and the locations of users, and obtained a 40% reduction in location uncertainty. Both CRPs and CRFs allow documents to be generated from a single mixture of topic, however they do not provide relationships between the topics. The nCRP remediates this by providing a tree-like structure in which the semantic is such that the parent topics are more general than the topic represented by their children. Document clustering and topic modelling can also be achieved using a combination of techniques such as the term-frequency inverse-document-frequency (tf-idf) metrics, and word embedding models. Online topic modelling captures and identifies emerging topics as streams of texts and their changes over time are received through online sources and are automatically clustered using LDA. A non-Markov online LDA Gibbs sampler topic model (OLDA) that enabled the current model to use new data [32] to guide the learning process of new generative process reflecting the dynamic nature of the data is employed and compared with LDA results of same datasets.

The authors in [33] evaluated several clustering techniques like word embedding, doc2vec, and tf-idf for documents and topic modelling using datasets from Twitter and Reddit and compared them with LDA. They observed that word embedding may be used effectively as a basis for document clustering, and that doc2vec and tf-idf weighted mean word embedding representation delivered better results than the average of word embedding vectors in document clustering tasks. The tf-idf method and LDA only gave results that were comparable to word2vec on data sizes with characters greater than 500. Automatic document summarization is another application of topic modelling, whose evaluation when compared to human manual summarization is difficult to achieve. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used in [34] to compare with human-annotated summarization as against that made with four semantic models (LSA, LDA, Word2Vec and Doc2Vec) and one frequency-based model (tf-idf) for document feature extraction. Another variant of LDA topic modelling is in the replacement of LDA’s parametrization of opaque topic types with categorical distributions of multivariate Gaussian distributions on embedding space [35]. Here, there is the assumption that documents, rather than consisting of sequence of word types, consist of word embedding sequences. These words in Gaussian LDA are ranked based on density assigned to them by the posterior predictive distribution.

Deep neural networks (DNN) have been successfully applied in the areas of image processing, speech and face recognitions. Neural network-based systems have also proved efficient in natural language processing (NLP). DNN has also found application in word level modelling by being able to predict word sequences from history of words. A hierarchical Recurrent Neural Network (HRNN) model [36] offered a two-step training approach for the approximation of sentence-level and word-level language model which converges in a pipeline style.

Modelling topics in a document on the social media is a trend that is gaining popularity in order to understand the similarity of topics of discussion among different online users, joining it with determining the community relationship among authors of same type of topic and unifying them using Topic-Link LDA algorithm is the attempt made in [37]. This seeks to understand the tie between not only topics but also between authors. In contrast to the traditional LDA topic modelling, an information filtering model, maximum matched pattern-based topic model seeks to model users’ interest in multiple topics rather than a single one in a document [38]. Maximum matched patterns are used to efficiently represent and rank documents.

III. MATERIALS AND METHODS

In pre-processing of data, it must be observed that documents are a mixture of topics which have some semantic coherency one with another. The standard human evaluation of topic modelling is designed to assess the performance of this semantic coherence among the topics. In the following section, we take corpora of documents and subject them to several topic modelling techniques in order to discover some observed topics, as well as determine the coherence among topics. Documents retrieved from Twitter and Kaggle machine learning repositories are subjected to training using different topic modelling algorithms and evaluated.

1. *Methodologies for Topic Modelling*

Three topic modelling techniques were employed and compared. They are the LDA, LSA/LSI, and HDP. The reason for employing these three techniques was to find out which one gives the best result and models the topics better using perplexity and coherency as a means of evaluation. Unsupervised machine learning approach was followed in developing these models. Two data embedding techniques are also employed and compared to determine if data embedding technique would have a significant impact on the topic models. The data embedding techniques used are bag of words (BOW), and the term-frequency inverse-document-frequency (tf-idf) metrics, word embedding models.

Gensim and Mallet frameworks for Python are used in Jupyter NoteBook for Python 3 as the program development environment. R Studio was used to retrieve live tweets. In the following section, descriptions of these experiments are provided.

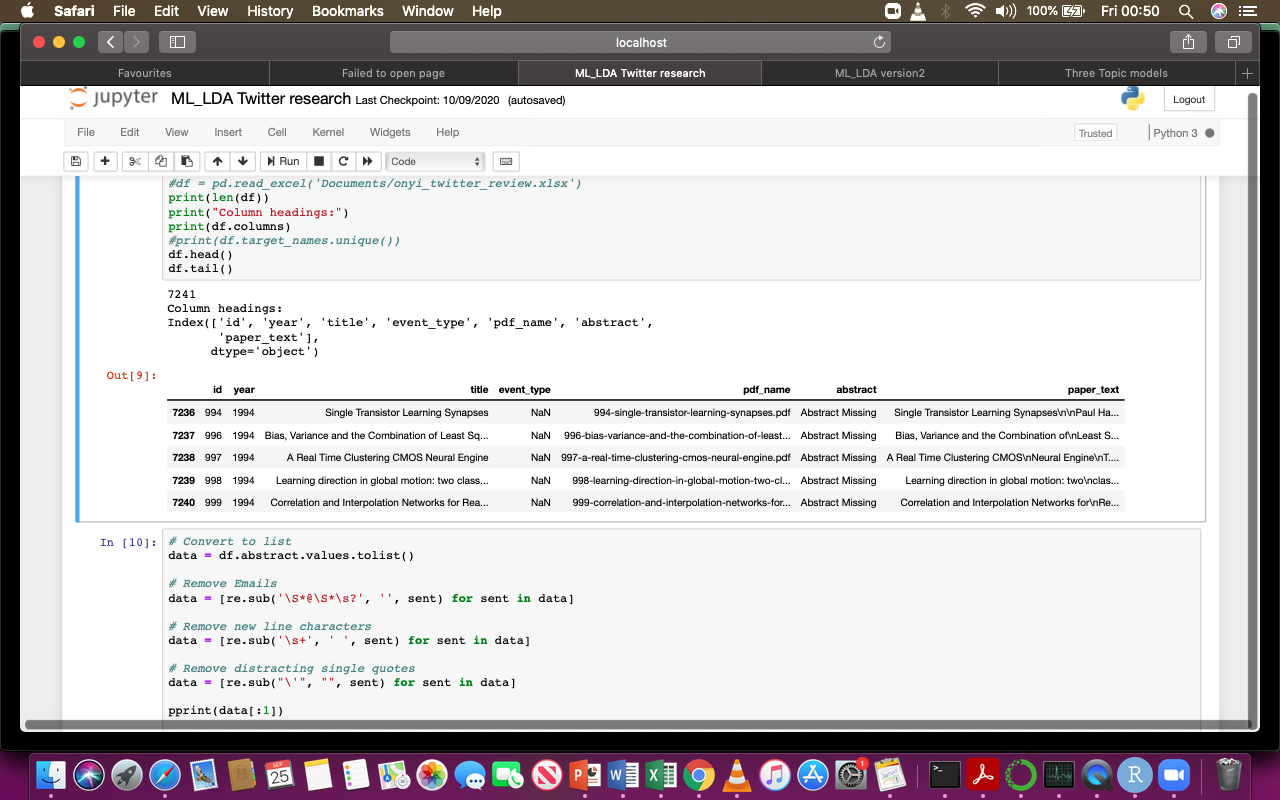
1. **Data sources**

Two major sources of data used for the research were from Kaggle and user tweets. The experiments were conducted using Neural Information Processing Systems (NIPS) conference papers between 1975 and 2017, a total of 7,241 articles downloaded from the Kaggle machine learning datasets (see fig. 4). NIPS offers one of the top global machine learning conferences whose topics range from deep learning and computer vision to cognitive science, as well as reinforcement learning [39]. The dataset contains the title, authors, and abstracts for all the NIPS papers within the date ranging from 1975 to 2017 amounting to 7,241 documents. The other dataset is user tweets on machine learning and AI. It contained 35,860 unorganized datasets (tweets) from 20 top AI and machine learning related handles retrieved from Twitter handles using Twitter API keys in R programming (see Table 1 and fig. 5, respectively).

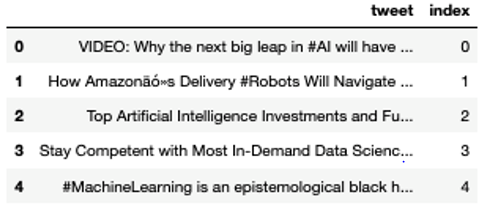
TABLE I

TWITTER HANDLES FOR DATA SOURCES

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SN** | **HANDLE** | **NAME OF ONWER** | **FOLLOWERS** | **FOLLOWING** | **rf** | **SPECIAL INTEREST** | **TOTAL NO. OF TWEETS** | **NO. OF TWEETS RETRIEVED** |
| 1 | @ylecun | **Yann LeCun** | 238.5k | 380 |  | optical character recognition and computer vision using convolutional neural networks (CNN). | 6,070 | 652 |
| 2 | @analyticsindiam | Analytics India Magazine | 12.2k | 353 |  | AI/ML enthusiasts Community in India | 17.9k | 1,798 |
| 3 | @andrewyng | Andrew Ng | 552.7k | 540 |  | Natural language processing, deep learning | 1,324 | 447 |
| 4 | @karpathy | Andrej Karpathy | 232.7k | 605 |  | Tesla’s neural networks autopilot specialist | 6687 | 3959 |
| 5 | @deeplearn007 | Imtiaz Adam | 33.6k | 118.3k |  | Insights on AI technologies | 125.9k | 7,355 |
| 6 | @sallyeaves | Prof Sally Eaves | 115.3k | 99.2k |  | Blockchain, AI, Cloud computing | 114.4k | 10,000 |
| 7 | @kirkdborne | Dr. Kirk Borne | 273k | 9,510 |  | Big data, data science, AI. | 129.9k | 8,863 |
| 8 | @paula\_piccard | Paula Piccard | 60.4k | 9,809 |  | Women in tech, digital transformation, AI, cyber security | 34.9k | 10,000 |
| 9 | @ronald\_vanloon | Ronald Vanloon | 237.6k | 180.7k |  | AI, big data, IoT, machine learning, analytics. | 93.5k | 10,000 |
| 10. | @spirosmargaris | Spiros Margaris | 106.2k | 15.7k |  | Fintech, Artificial Intelligence (AI) and Blockchain | 208.8k | 15,000 |
| 11 | @drfeifei | Prof. Fei Fei | 1,519 | 469 |  | Improving human condition through AI Research | 374.4k | 516 |
| 12 | @tamaramccleary | Tamara McCleary | 306.8k | 207.1k |  | Social Media Analytics, Social Media Strategy and Influencer Programs. | 108.6k | 3,606 |
| 13 | @EvanKirstel | Evan Kirstel | 296.4k | 273.2k |  | Social media marketing, IoT, cloud, Telecom, 5G | 988.7k | 11,478 |
| 14 | @MikeQuindazzi | Mike Quindazzi | 156.9k | 3,134 |  | Development in AI and Robotics | 54.7k | 11,169 |
| 15 | @hmason | Hillary Mason | 126.7k | 1,870 |  | Robotics and intelligence research, Data science in Residence | 19.4k | 64 |
| 16 | @antgrasso | Anthonio Grasso | 179k | 44.9k |  | Digital Transformation, Mobile DevOps, Digital Tokens and AI Marketing | 41.7k | 15,000 |
| 17 | @fisher85m | Michael Fisher | 89k | 11.3k |  | Cyber security and AI innovation | 32.4k | 5,697 |
| 18 | @andy\_fitze | The AI Andy Fitze | 90.7k | 41.2k |  | Expert predictions for 2020’s AI development | 4,790 | 331 |
| 19 | @jblefevre60 | Jean-Baptiste Lefevre | 80.2k | 5,551 |  | Innovations in AI, ML, cloud | 183.9 | 15,000 |
| 20 | @PierrePinna | Pinna Pierre | 53.1k | 50.1k |  | Innovations and new technology in Robotics in French | 79k | 4,989 |
|  |  |  |  |  |  |  | **Total** | **135,860** |



**Fig. 4.** Structured document (NIPS papers)



**Fig. 5.** Unstructured document (tweets)

1. **Procedure**

The Gensim model in Python is used for the topic modelling experiments on Twitter raw dataset and the NIPS conference papers. In both experiments the procedure is described as follows:

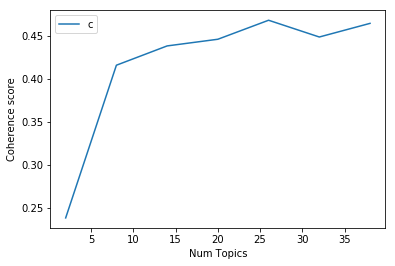
1. **Loading data**: all data files were saved as comma separated values (csv) and read into the programs using the Pandas library. Fig 4 and fig. 5 show the head sections of the two datasets used in the two respective experiments as read by Pandas.
2. **Data Cleaning:** Stop words remover from the Natural Language Tool Kit (NLTK) is used to remove words that do not add any specific meaning to the documents. The Spacy en (English) model is used for text pre-processing which includes lemmatization. Lemmatization helps to convert words to their root word, for example ‘walking’ is lemmatized to ‘walk’, ‘machines’ to ‘machine’, etc.
3. **Phrase Modelling:** Bi-grams and Tri-grams: some words appear as two or three words together frequently in a document. It is important to recognize such words. The Gensim’s Phrase model is used to build bigrams and trigrams.
4. **Data transformation:** Corpus and Dictionary. We created dictionary and corpus required for Topic Modelling. The two main inputs to the LDA topic model are the dictionary and the corpus. Gensim creates a unique id for each word in the document.

Other procedures include Base Model Performance, Hyper-parameter Tuning, Final Model, and Visualization of Results.

1. *How to Determine the Optimal Number of Topics*

We created a function that determined the coherence values of different numbers of topics and plotted them. This function trains multiple models and prints their respective coherence values. The topic with the highest coherence value before the curve begins to flatten is usually the best one to choose. In this case we chose topic number of 26 as shown in the fig. 6. This topic has index position 4. So we are more informed now, so we repeat the experiment with 26 topics.

The Python program for topic models was run repeatedly for the LDA, LSI and HDP models across topic numbers 2, 8, 14, 20, 26, 32, 38, 44, and 50 using both data sets one after the other, as well as interchanging the word embedding. The results are reported in subsequent sections. Using varying topic numbers helps us to have true picture of the topic models performances rather than base our judgement on a single point. The results are shown using tables and visualizations.



**Fig. 6.** Plot for determining number of topics

1. *Topic Modelling Evaluation Metrics*

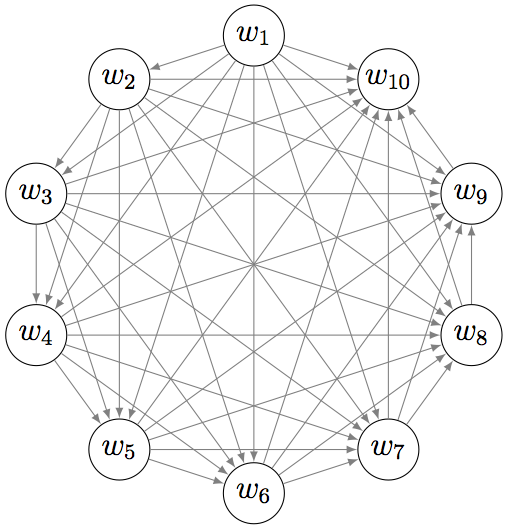
LDA topic modelling like any other unsupervised machine learning algorithm is hard to evaluate since there are no hard truths against them. There are many methods for evaluating topic modelling performance. These include coherence measure [40-42], perplexity [43], Chib-style estimator and “left-to-right” algorithm [44], and visualization methods. Two popular methods for evaluating the performance of topic models are coherence and perplexity. Coherence measure, an average of pairwise word similarities formed by top words in a given topic has been adopted [40] for measuring how good or bad a topic is. This is because topic models do not guaranty the interpretability of their results. There are two types of coherence measures, intrinsic and extrinsic [41]. Both compute the sum of pairwise scores on words w1,..,wn (see fig. 7) used in explaining the topic as in equation 2, for describing the top n words by frequency in topic p(w|k)p(w|k). This measure is seen as the sum of all the edges of the graph in fig 7.

Coherence = (2)

1. **Intrinsic (Umass)**: this measure compares a word only to the previous and next words, respectively. The words have to be an ordered set for this to work. It uses a pairwise score function, an empirical conditional log-probability with smoothing count by adding one to D(wi,wj) in order to avoid calculating the log of zero. Where p(w) is the probability of seeing word, wi in a random document, and p(wi,wj) is the probability of both words occurring in a random document.

The Umass score is given as ScoreUMASS(wi,wj)= (3)

### **Extrinsic Measure (UCI)**: uses a pointwise score function known as Pointwise Mutual Information (PIM) in which single words are paired with each other. The UCI score is given in (4). ScoreUCI(wi,wj)= (4)



**Fig. 7.** Word connections in a topic (source: Pleple [41])

Topic Coherence measure is a good way to compare different topic models based on their human-interpretability.

1. **Perplexity score**: Log-likelihood of held-out test data is the most common way to evaluate a probabilistic model [45] such as the LDA. Like in supervised learning, the output or the trained dataset in LDA can be split into training and test sets, respectively. In LDA, a test set is a collection of unseen documents, wd, and q describes the model topic matrix, while a, the hyperparameter represents the topic distribution for the documents. The log-likelihood is evaluated as in (4). A higher log-likelihood implies a better model.

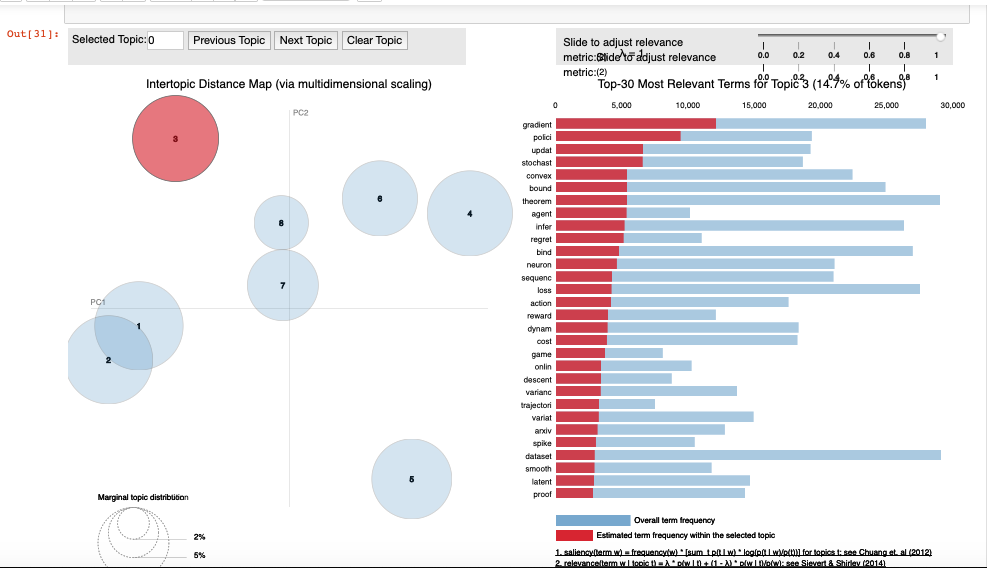
L(w) = logp(w|q,a)= (4)

Perplexity is the traditional method used for held-out documents Wd in a topic model, and it is given as in (5). Perplexity measures the ability of a model to generalize and predict new text documents [43]. It indicts the number of equally likely words to occur in a document at an arbitrary position. The LDA model uses the log\_perplexity method which takes a bag of words corpus as input and returns the perplexity value.

Perplexity(test\_set w) = (5)

The lower the perplexity value the better the model [43,45].

1. **Visual Evaluation**: The LDA visualization tool pyLDAvis, is another way to evaluate the result of topic models. It is a Python interactive library for model visualization that represents topics extracted from the LDA model in form circles or bubbles. The bigger the circle the higher the percentage of number words that make up the topic. The number of bubbles is determined by the number of topics we want the LDA model to generate for us (see fig. 8). The bubbles are numbered serially according to the number of topics. As we move from one bubble to another the topics change dynamically, displaying the words that make up each topic, as well as the overall frequency of each word in the corpus represented by blue bars, and the estimated number of times a term is generated by a given topic shown with red bar. For instance, the word gradient appeared in about 28,000 places, but was used about 12,000 times in topic 3. The further away the bubbles are from each other the less similar the topics are, while the closer they are, the more similar the topics. Judging by the terms in each topic, one can conclude on what the topic is about. By default the pyLDAvis shows the top 30 most relevant terms in a topic.



**Fig. 8.** pyLDAvis visualization of topic models

IV. RESULTS AND DISCUSSIONS

In this section we present the methods of evaluating topic models as well as the results of the following experiments; (i) a comparison of the three topic modelling techniques; LDA, LSI, and HDP. The performances are compared to determine the best algorithm. (ii) Determining a better embedding technique between the bag of words (BOW) embedding, and the Term Frequency Inverse Document Frequency (TF-IDF) for topic modelling. (iii) The section also presents the results of the comparison of Gensim and Mallet Models.

1. *Results of Experiments*
2. **Unstructured vs Structured Datasets**

Table 2 shows the summary of LDA evaluation results for the two datasets, the unstructured tweets and the structured NIPS datasets.

TABLE II

EVALUATION RESULTS FOR THE TWO DATASETS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | No of documents | Coherence Umass\_Score | Coherence UCI\_Score (c\_v) | Perplexity |
| Live tweets | 135,924 | -13.53022 | 0.3945671 | -22.73715 |
| NIPS papers | 7,241 | -0.72836806 | 0.457887571 | -8.688008 |

In coherency the NIPS dataset had better performance. This is understandable since the dataset is structured, it is easier for the algorithm to organize the topics better, but the raw tweets had better perplexity performance, though some authors argue that perplexity is not a good measure of performance. The NIPS dataset performed 86% better than the raw tweets using the UCI coherence score and 5.4% better than the tweets using the U\_mass coherence score.

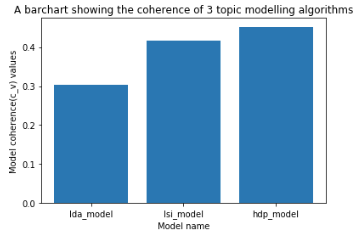
1. **BOW Topics Results for LDA, LSI, HDP on Tweets**

TABLE III

A COMPARISON OF THE PERFORMANCES OF THE THREE MODELS ON THE TWO DATASETS

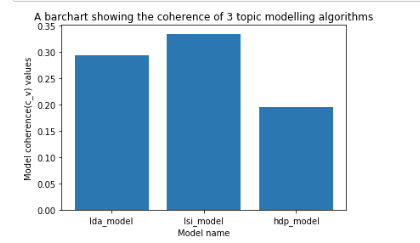
|  |  |  |
| --- | --- | --- |
| **Model** | **NIPS Coherence** | **Tweets Coherence** |
| LDA | Perplexity: -7.570  0.3026072675 | Perplexity: -7.0080  0.29372262845 |
| LSI | 0.4166050449 | 0.33385338957 |
| HDP | 0.45218042459 | 0.195466477917 |

All the models performed better in structured NIPS datasets than in the unstructured tweets, agreeing with the general findings in Table 2. Both the coherency and the LDA perplexity attests to this. The HDP model performed better than the LSI and LDA models in that order using the NIPS dataset, as shown in fig. 9.



**Fig. 9.** NIPS coherence values for LDA, LSI, and HDP Models

The LSI beat the LDA and the HDP in that order using the tweets dataset, see fig. 10.



**Fig. 10.** Tweets coherence values for LDA, LSI, and HDP Model

TABLE IV

EVALUATION SCORES ACROSS THE THREE MODELS AND TOPICS IN TWEETS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No\_topics** | **LDA\_Perplexity** | **LDA\_CV** | **LSI\_CV** | **HDP\_CV** |
| **2** | -6.992287979 | 0.26038149399 | 0.2109477416 | 0.1971698105 |
| **8** | -7.008076761 | 0.29372262845 | 0.3338533896 | 0.1954664779 |
| **14** | -9.042080899 | 0.35260845743 | 0.3024220818 | 0.1964664779 |
| **20** | -11.41537552 | 0.39173503481 | 0.3092057430 | 0.1948254684 |
| **26** | -12.62313976 | 0.42963160923 | 0.2963431871 | 0.2011519489 |
| **32** | -13.85463406 | 0.43276902837 | 0.3158016858 | 0.1954439597 |
| **38** | -15.02677988 | 0.45899579357 | 0.3229799664 | 0.1957414457 |
| **44** | -16.25851165 | 0.46683228728 | 0.3325065428 | 0.1983655346 |
| **50** | -17.91880002 | 0.46980153149 | 0.3405411231 | 0.1965291292 |

Table 4 and fig. 11 show the results of the evaluation performances of the LDA, LSI and HDP models across topic numbers 2, 8, 14, 20, 26, 32, 38, 44, and 50. Coherence scores increases consistently as the number of topics increases in the LDA model, while perplexity decreases as topic increases. Both the LSI and HDP have fluctuating results. Though the HDP does not have a method for number of topics, each run of the program produces a different coherence value as we varied the number of topics in LDA and LSI.

**Fig. 11.** Coherence scores of LDA, LSI and HDP against topic number

Unlike the result in section 4.2.1, where the winner was inconclusive, the result in Table 4 shows that the LDA was a better model in terms of consistency with number of topics and in having the highest coherence values in all but one case, that is when the number of topics is 8. LSI ranked second, and HDP performs rather poorly.

1. *Bag of Words vs Term Document Inverse Frequency*

In order to find out the better embedding method between the two most frequently used corpora, performance evaluation was done by classifying the NIPS test dataset documents using LDA Bag of Words model and LDA TF-IDF model. The following results were obtained and compared in Table 5.

TABLE V

RESULT TOPICS EVALUATION OF LDA-BOW AND LDA-TF-IDF EMBEDDING

|  |  |  |
| --- | --- | --- |
|  | Perplexity | Coherence |
| LDA-BOW | -7.739107959 | 0.49215338 |
| LDA-TFDIF | -7.811492707 | 0.25257377 |

The BoW embedding out-performed the TF-IDF by 50.9% judging by the coherency values, and also gives a better result in perplexity evaluation. When used to predict unseen documents, the BoW also performed better in classifying the document into topics having a score of 0.99 (99%) accuracy, while TD-IDF has an accuracy score of 0.58 (58%), as the best results, respectively.

1. *Results from Twitter Data: Gensim vs Mallet Models*

The table 6 compares the results from the Twitter dataset as observed from the Gensim and Mallet models, respectively in a-20 topic LDA model. Selected results were based on term weights from 0.2 or 20% and above term probabilities in each topic. The Mallet had 13 terms while the Gensim had 9 terms whose values are 20 percent and above. Both Mallet and Gensims have 4 topics in common with varying weights. The words are: ai, developer, machine, and machinelearne are the common words. The Mallet model outperforms the Gensim in its ability to identify more number of topics and creating more relevant topics.

TABLE VI

GENSIM VS MALLET MODELS

|  |  |  |
| --- | --- | --- |
| TOPICS | MALLET | GENSIM |
| 0 | 0.276\*"deeplearne |  |
| 1 | 0.277\*"machinelearne |  |
| 2 | 0.224\*"machinelearne" |  |
| 3 | 0.289\*"datascience | 0.352\*"datum |
| 4 | 0.321\*"https" |  |
| 5 | 0.247\*"machinelearne |  |
| 6 |  | 0.221\*"create |
| 8 | '0.557\*"ai" |  |
| 9 | 0.218\*"developer" | 0.290\*"python |
| 10 | '0.339\*"machine" |  |
| 12 | 0.245\*"twitter |  |
| 13 |  | 0.284\*"developer" 0.242\*"complete |
| 14 |  | 0.300\*"machine" 0.204\*"learn |
| 16 | 0.259\*code 0.214\*programming 0.206\*programmer |  |
| 18 |  | 0.391\*"ai" 0.320\*"machinelearne |

1. *Measuring Twitter Influence on AI and ML Research*

Comparing the Correlation between the Tweets and the NIPS Papers.

In this section we want to find out if the tweets could have effects on the direction of research in the fields of AI and ML. The publications in the NIPS papers tell us the direction of research in the fields of AI and ML. Could what people tweet about have any influence on what researchers would want to work on? Do we have means of measuring this influence? This is what this section aims at addressing.

In addressing these questions, the first point of call is to look at the people who tweet on these subjects and find ways to measure their influence in the social media, viz-a-viz, the research community. Next, we perform a commonality analysis between the tweets and the NIPS papers.

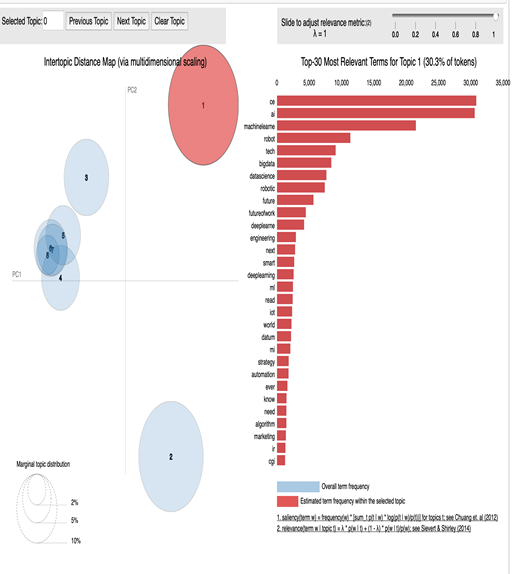
1. **Twitter Influences and Influencers**

Does tweets influence research direction in the AI/ML ecosystem? If it does, who are the influencers?

Fig. 12 is a pyLDAvis visualization of Twitter topic models for 8 topics. Moving your cursor over the bubbles (twitter topics) gives you idea on what the topic of discussion is. For example, topic 1 (30.2% of tokens) discusses AI, machine learning, robotics, data science, etc. This shows that the topic is about AI and its associated subfields. Topic 2 (25.6% of tokens) discuses innovation in AI such as in research, gaming, invention, prediction, etc. Topic 3 (12.3% of tokens) discusses fintech; top trending fintechs which include fintech companies, software prediction models, etc. Topics 4-8 overlap at varying degrees and they are discussing machine learning technologies and applications in the areas of skills development, health, etc.

With over 187 million active twitter users as at the third quarter of 2020, only a small percentage of this number are influencers, or what is called the alpha group [46]. These group of people are characterized by the large amount of followership and by the contents they generate, which usually get distributed through retweets and mentions. These contents may even be distributed in other microblogs to people who may not be their direct followers. This widespread contacts gives the influencer more influence over their user reach. Some other researchers have other names for describing influencers, such as opinion leaders, innovators, prestigious or authoritative actors [47]. In some other cases they may be regarded as topical experts in specific domains, such as machine learning and artificial intelligence as shown in Table 1.

Riquelme and Gonzalez-Cantergiani [47] provided a table of Twitter metrics for measuring popularity, influence, activity, etc. For instance in measuring a user’s popularity the formula F1/F3 or Twitter Follower-Followee ratio (TFF) is used, where F1 is the number of followers and F3 is the number of followee. In Table 7 we show a modified version of Table 1 after calculating the users’ popularity and ranking the top 10 influencers in the subject of discussion.



**Fig. 12.** pyLDAvis visualization of Twitter topic models for 8 topics

TABLE VII

POPULAR TWITTER AL AND ML USERS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/N** | **HANDLE** | **NAME OF OWNER** | **TFF** (F1/F3) | **Rank** (F1/(F1+F3)) |
| 1 | @andrewyng | Andrew Ng | 1023.518519 | 0.999023932 |
| 2 | @ylecun | **Yann LeCun** | 627.6315789 | 0.998409243 |
| 3 | @karpathy | Andrej Karpathy | 384.6280992 | 0.997406828 |
| 4 | @hmason | Hillary Mason | 67.7540107 | 0.985455394 |
| 5 | @MikeQuindazzi | Mike Quindazzi | 50.06381621 | 0.980416661 |
| 6 | @analyticsindiam | Analytics India Magazine | 34.56090652 | 0.971879232 |
| 7 | @kirkdborne | Dr. Kirk Borne | 28.70662461 | 0.966337475 |
| 8 | @jblefevre60 | Jean-Baptiste Lefevre | 14.44784723 | 0.935266061 |
| 9 | @fisher85m | Michael Fisher | 7.876106195 | 0.887337986 |
| 10 | @spirosmargaris | Spiros Margaris | 6.76433121 | 0.871205906 |

The authors in [46] observed that the influence of a tweet can be measured in terms of the number of users affected by the message in the tweet. This influence can be in terms of intensity, cognitive, emotional, and conductual impact. Our study further examines the influence of the top 5 influencers in table 7 with respect to their H-index, citations, etc. The result is shown in table 8.

TABLE VIII

TOP 5 TWITTER INFLUENCERS ON GOOGLE SCHOLAR AND SCOPUS (ACCESSED ON: 01-03-2021)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **GOOGLE SCHOLAR** | | | **SCOPUS** | | |
| **S/N** | **TWITTER HANDLE** | **NAME** | **H-INDEX** | **I10-INDEX** | **NO. OF CITATION** | **H-INDEX** | **DOCUMENTS** | **CITATION** |
| 1. | @andrewyng | Andrew Ng | 130 | 279 | 166833 | 88 | 203 | 65077 |
| 2. | @ylecun | Yann LeCun | 126 | 294 | 186756 | 74 | 192 | 72556 |
| 3. | @karpathy | Andrej Karpathy | 11 | 11 | 35093 | 12 | 13 | 18298 |
| 4. | @hmason | Hilary Mason | N/A | N/A | N/A | N/A | N/A | N/A |
| 5. | @mikequindazzi | Mike Quindazzi | N/A | N/A | N/A | N/A | N/A | N/A |

Table 8 shows the top five twitter influencers, sourced from Google Scholar and Scopus websites.

1. Andrew Ng is the cofounder of Coursera and an adjunct professor in Computer Science Department of Stanford University. He leads the Stanford Machine Learning group.
2. Yann LeCun is the Chief A.I Scientist at Facebook and a silver professor at the Courant Institute, New York University.
3. Andrej Karpathy is the director of A.I. and Autopilot Vision at Tesla. He specializes in deep learning and computer vision. He was previously a research scientist at Stanford University.
4. Hilary Mason is the cofounder of @hiddendoor co. and former founder of Fastforward labs.
5. Mike Quindazzi is the sales leader at Digital Alliance.
6. **Top Five Machine Learning Researchers (By Research Interest) on Google Scholar**
7. Geoffrey Hinton (citations: 425,125): He is an Emeritus Professor in the University of Toronto.
8. Yoshua Bengio (citations: 376,447): A professor of computer science from the University of Montreal.
9. Robert Tibshirani (citations: 367,517): A professor of Biomedical data science, Stanford University.
10. Trevor Hastie (citations: 255,606): Professor of Statistics, Stanford University.
11. Andrew Zisserman (citations: 244,410): is a professor of computer science in the University of Oxford and a computer vision researcher.
12. **Findings**

The top five machine learning influencers identified on twitter (see Table 8) are mostly industry-based professionals, with some combining research and academia with industry experience.

When compared with the data from google scholar, it was observed that the top five machine learning researchers with a google profile have greater citations than those of the top five twitter influencers.

Most of the top researchers on google scholars who are professors and top researchers in the field of machine learning are not very active on Twitter, though they have Twitter profiles. It is reasonable to infer that the top AI/ML influencers in table 7 are able to exert this influence because of their activeness on Twitter unlike the more cited scholars who are less active on the social media.

Since our findings show that the LDA model out-performed the other two models, we proceeded to compare the LDA evaluation metrics across the NIPS papers and the user tweets. Table 9 shows the results of the coherence and perplexity values from varying numbers of topics (topics 2-50). Applying t-test to determine if there is a significant difference between the results of the two datasets under study.

TABLE IX

EVALUATION SCORES ACROSS OF THE LDA MODEL AND TOPICS IN THE 30 YEARS NIPS PAPERS AND TWEETS.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. of Topics | LDA\_CV for NIPS Papers | LDA\_CV for AI\_ML Tweets | LDA\_Perplexity  (NIPS Papers) | LDA\_Perplexity  (Tweets) |
| 2 | 0.333050244307 | 0.260381 | -7.738696634 | -6.992287979 |
| 8 | 0.412244281318 | 0.293723 | -7.829689027 | -7.008076761 |
| 14 | 0.465193850187 | 0.352608 | -8.785319047 | -9.042080899 |
| 20 | 0.467526397358 | 0.391735 | -9.389655864 | -11.41537552 |
| 26 | 0.476460525678 | 0.429632 | -9.899802937 | -12.62313976 |
| 32 | 0.46242261588 | 0.4327690 | -10.4234657 | -13.85463406 |
| 38 | 0.47927547322 | 0.4589958 | -10.9348307 | -15.02677988 |
| 44 | 0.47782591896 | 0.4668323 | -11.4634772 | -16.25851165 |
| 50 | 0.46616613333 | 0.4698015 | -11.9916904 | -17.91880002 |

In order to test the results we propose our hypotheses as follows:

Ho: There is no significant difference between the topics generated by the NIPS papers and those by user tweets.

H1: There is a significant difference between the results of the two datasets.

A t-test returned a t-statistic value of 1.7445, a p-value of 0.09873 at 95% confidence interval for the difference in means from -0.01150148 to 0.11898767 for the NIPS papers and user tweets datasets, respectively. The mean in groups NIPS paper and user tweets returned 0.4489073 and 0.3951642, respectively.

With a p-value greater than 0.05, the null hypotheses (Ho) is accepted and the alternative hypotheses (H1) is rejected. This means that there is no significant difference between the results obtained using the NIPS papers and the user tweets. The same result is true when we consider the perplexity scores, which returned a p-value of 0.1175, which is also not statistically significant.

V. CONCLUSION

This study set out to provide methods for online data collection and analytics by systematically showing easy steps to researchers who are challenged by non-acceptance of their research papers due to either lack of knowledge on how to mine research data from the web or how and what tools to apply for insightful data analytics. The study used a-40 year NIPS conference papers dataset and Twitter users tweets to gauge the amount of influence tweets could have on the direction of academic research in the areas of artificial intelligence and machine learning. Three topic modeling algorithms were applied, the LDA, LSI, and HDP. The algorithms successfully identified AI/ML topics at varying degrees. Their results were compared across different number of topics, and the LDA came out with the best results, which is consistent with literature. Furthermore, embedding methods were also compared within the LDA model, the Bag of words and the tf-idf, respectively. The bag of words being the traditional embedding method for LDA came out better in discovering more topics within the corpora. When the evaluation metrics of the LDA model were compared between the NIPS papers and the user tweets, it was discovered that the NIPS dataset, an organized dataset showed better performance. However, a t-test showed that the difference between the results of the two datasets was not significant. The author in [47], observed that one problem of Twitter measures based on LDA is that the traditional LDA algorithms are used for larger texts than tweets, so it is necessary to use specific variation. One of the striking findings of this paper is the ability of the research to identify the top influencers in AI/ML research from tweets using topic modeling. The top five machine learning influencers identified on Twitter are mostly industry-based professionals, some combining research and academia with industry experience. These influencers are able to exert this influence because of their activeness on Twitter unlike the more cited scholars who are less active on social media.

References

[1] B. Batrinca and P.C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” AI & Soc, 30, 89–116, 2015, <https://doi.org/10.1007/s00146-014-0549-4>.

1. L. Bornmann and R. Haunschild, “T factor: A metric for measuring impact on Twitter,” Malaysian Journal of Library & Information Science, Vol. 21, no. 2, 13-20, 2016.
2. J. Ortega, The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact (citations),” Aslib Proceedings, 69, 10.1108/AJIM-02-2017-0055, 2017.
3. S. Haustein, “Scholarly Twitter metrics, Handbook of Quantitative Science and Technology Research,” 2018.
4. T. Hofmann, Probabilistic Latent Semantic Indexing, In SIGIR’99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57, ACM Press, 1999.
5. D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, 3, 993–1022, 2003.
6. B. Grün and K. Hornik, “Topicmodels: An R Package for Fitting Topic Models,” Journal of Statistical Software, 40 (13), pp. 1-30, ISSN 1548-7660, 2011.
7. R. Nagar, Q. Yuan, C. C. Freifeld, M. Santillana, A. Nojima, R. Chunara and J.S. Brownstein, “A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives,” Journal of medical Internet research, 16(10), e236, 2014.
8. A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth and E. J. Langer, “Forecasting the onset and course of mental illness with Twitter data,” Scientific reports, 7(1), 1-11, 2017.
9. A. Ashok, M. Guruprasad, C. O. Prakash and S. S. Shylaja, “A Machine Learning Approach for Disease Surveillance and Visualization using Twitter Data,” In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), IEEE, pp. 1-6, 2019.
10. C. Allen, M. H. Tsou, A. Aslam, A. Nagel, J. M. Gawron, “Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza” PloS one, 11(7), e0157734, 2016.
11. K. Lee, A. Agrawal and A. Choudhary, “Real-time disease surveillance using twitter data: demonstration on flu and cancer,” In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1474-1477, 2013.
12. G. Caldarelli, A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni and G. Riotta, “A multi-level geographical study of Italian political elections from Twitter data,” PloS one, 9(5), e95809, 2014.
13. M. Anjaria and R. M. R. Guddeti, “Influence factor based opinion mining of Twitter data using supervised learning,” In 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), IEEE, pp. 1-8, 2014.
14. D. Gayo-Avello, P. T. Metaxas and E. Mustafaraj, “Limits of electoral predictions using twitter,” In Proceedings of the fifth international AAAI conference on weblogs and social media. Association for the Advancement of Artificial Intelligence, 2011.
15. H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan, “A system for real-time twitter sentiment analysis of 2012 us presidential election cycle,” In Proceedings of the ACL 2012 system demonstrations, pp. 115-120, 2012.
16. L. Wang and J. Q. Gan, “Prediction of the 2017 French election based on Twitter data analysis,” In 9th Computer Science and Electronic Engineering (CEEC), IEEE, pp. 89-93, 2017.
17. M. Choy, M. L. Cheong, M. N. Laik and K. P. Shung, “A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction,” arXiv preprint arXiv:1108.5520, 2011.
18. M. Skuza and A. Romanowski, “Sentiment analysis of Twitter data within big data distributed environment for stock prediction,” In 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, pp. 1349-1354, 2015.
19. Y. Mao, W. Wei, B. Wang and B. Liu, Correlating S&P 500 stocks with Twitter data, In Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research, pp. 69-72, 2012.
20. V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, “Sentiment analysis of Twitter data for predicting stock market movements,” In 2016 international conference on signal processing, communication, power and embedded system (SCOPES), IEEE, pp. 1345-1350, 2016.
21. X. Wang, M. S. Gerber and D. E. Brown, “Automatic crime prediction using events extracted from twitter posts,” In International conference on social computing, behavioral-cultural modeling, and prediction, Springer, Berlin, Heidelberg, pp. 231-238, 2012.
22. Z. Miller, B. Dickinson, W. Deitrick, W. Hu and A. H. Wang, Twitter spammer detection using data stream clustering, Information Sciences, 260, 64-73, 2014.
23. S. Munir, S. Wasi and S. I. Jami, A Comparison of Topic Modelling Approaches for Urdu Text, Indian Journal of Science and Technology, 12, 45, 2019.
24. R. Vīksna, M. Kirikova and D. Kiopa, Exploring the Use of Topic Analysis in Latvian Legal Documents, 2018.
25. J. O'Neill, C. Robin, L. O'Brien and P. Buitelaar, An analysis of topic modelling for legislative texts, CEUR Workshop Proceedings, 2016.
26. C. P. George, D. Z. Wang, J. N. Wilson, L. M. Epstein, P. Garland and A. Suh, “A machine learning based topic exploration and categorization on surveys,” In 2012 11th International Conference on Machine Learning and Applications, Vol. 2, IEEE, pp. 7-12, 2012.
27. D. J. Hu, “Latent dirichlet allocation for text, images, and music” University of California, San Diego, 2009, Retrieved April, 26, 2013.
28. P. Xie, Y. Deng and E. Xing, “Diversifying restricted boltzmann machine for document modeling,” In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1315-1324, 2015.
29. I. Sato and H. Nakagawa, “Topic models with power-law using Pitman-Yor process,” In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 673-682, 2010.
30. A. Ahmed, L. Hong, A. Smola, “Nested chinese restaurant franchise process: Applications to user tracking and document modeling,” In International Conference on Machine Learning, pp. 1426-1434, 2013.
31. L. AlSumait, D. Barbará and C. Domeniconi, “On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking,” In 2008 eighth IEEE international conference on data mining,” IEEE, pp. 3-12, 2008.
32. S. A. Curiskis, B. Drake, T. R. Osborn and P. J. Kennedy, An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit, Information Processing & Management, 57(2), 102034, 2020.
33. M. Campr and K. Ježek, Comparing semantic models for evaluating automatic document summarization, In International Conference on Text, Speech, and Dialogue, Springer, Cham. pp. 252-260, 2015.
34. R. Das, M. Zaheer and C. Dyer, Gaussian lda for topic models with word embeddings, In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Volume 1: Long Papers, pp. 795-804, 2015.
35. R. Lin, S. Liu, M. Yang, M. Li, M. Zhou and S. Li, Hierarchical recurrent neural network for document modeling, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 899-907, 2015.
36. Y. Liu, A. Niculescu-Mizil and W. Gryc, “Topic-link LDA: joint models of topic and author community,” In proceedings of the 26th annual international conference on machine learning, pp. 665-672, 2009.
37. Y. Gao, Y. Xu and Y. Li, Pattern-based Topics for Document Modelling in Information Filtering. IEEE Transactions on Knowledge and Data Engineering, 27(6), 1629-1642, 2014, http://dx.doi.org/10.1109/TKDE.2014.2384497.
38. Kaggle: NIPS 2015 Papers. Available at: https://www.kaggle.com/benhamner/nips-2015-papers.
39. F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both, “Evaluating topic coherence measures,” arXiv preprint, arXiv:1403.6397, 2014.
40. Q. Pleple, Topic Coherence to Evaluate Topic Models, 2013, Available at: http://qpleple.com/topic-coherence-to-evaluate-topic-models, Retrieved: 27th January, 2021.
41. M. Röder, A. Both and A. Hinneburg, “Exploring the space of topic coherence measures” In Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399-408, 2015.
42. A. De Waal and E. Barnard, Evaluating topic models with stability, Available at: <http://researchspace.csir.co.za/dspace/bitstream/handle/10204/3016/de%20Waal1_2008.pdf?sequence=1&isAllowed=y> , 2008.
43. H. M. Wallach, I. Murray, R. Salakhutdinov and D. Mimno, “Evaluation methods for topic models,” In Proceedings of the 26th annual international conference on machine learning, pp. 1105-1112, 2009.
44. Q. Pleple, “Perplexity to Evaluate Topic Models,” Available at: http://qpleple.com/perplexity-to-evaluate-topic-models/ (2013), Accessed 28th January, 2021.
45. I. Anger and C. Kittl, “Measuring influence on Twitter,” Conference: Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, Available from: https://www.researchgate.net/publication/220866465\_Measuring\_influence\_on\_Twitter, 2011, Accessed 15 February 2021, DOI: 10.1145/2024288.2024326.
46. F. Riquelme and P. Gonzalez-Cantergiani, “Measuring user influence on Twitter: A survey,” Journal of Information Processing and Management, 52(5), Available at: https://www.researchgate.net/publication/281427547\_Measuring\_user\_influence\_on\_Twitter\_A\_survey, 2016, Accessed 15 February 2021, DOI:10.1016/j.ipm.2016.04.003.

**Collins N. Udanor**, photograph and biography not available at the time of publication.

**Onyinye E. Nweke**, photograph and biography not available at the time of publication.

**George E. Okereke**, photograph and biography not available at the time of publication.

1. *(Corresponding author: Onyinye .E. Nweke).*

   The authors are with the Department of Computer Science, University of Nigeria, Nsukka Enugu State, Nigeria (e-mail: collins.udanor@unn.edu.ng; onyinye.nweke@unn.edu.ng; george.okereke@unn.edu.ng). [↑](#footnote-ref-1)