November 2020

Bibliometric Survey on Incremental Learning in Text Classification Algorithms for False Information Detection

Yashoda Narayanprasad Barve Mrs.
Suryadatta Group of Institutes, Pune Maharashtra, India, ynb994@gmail.com

Preeti Mulay Dr.
Symbiosis Institute of Technology, Pune, India, preeti.mulay@sit.edu.in

Bibliometric Survey on Incremental Learning in Text Classification Algorithms for False Information Detection

Yashoda Barve, Dr. Preeti Mulay

Bibliometric Survey on Incremental Learning in Text Classification Algorithms

Yashoda Barve[a], Dr. Preeti Mulay[b]

[a] Suryadatta College of Management Information Research &Technology, affiliated to Pune University, Pune, India
[b] Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed) University, Pune, India

ABSTRACT
The false information or misinformation over the web has severe effects on people, business and society as a whole. Therefore, detection of misinformation has become a topic of research among many researchers. Detecting misinformation of textual articles is directly connected to text classification problem. With the massive and dynamic generation of unstructured textual documents over the web, incremental learning in text classification has gained more popularity. This survey explores recent advancements in incremental learning in text classification and review the research publications of the area from Scopus, Web of Science, Google Scholar, and IEEE databases and perform quantitative analysis by using methods such as publication statistics, collaboration degree, research network analysis, and citation analysis. The contribution of this study in incremental learning in text classification provides researchers insights on the latest status of the research through literature survey, and helps the researchers to know the various applications and the techniques used recently in the field.

KEYWORDS
Bibliometric Analysis, Incremental Learning, False information, Text Mining, Text Classification, Machine Learning

Introduction

Text mining has become an active research area and gained popularity among researchers particularly in text classification to find relevant information from the documents. Text classification aims at identifying, understanding and categorizing or tagging the text documents based on the content of its keywords to get structured data (Al-Diabat 2012; Guo et al., 2010; Lakshmi Prasanna and Rajeswara Rao, 2017). Most of the text classification systems involve extracting features, dimensionality reductions, selection of classifier and evaluations (Kowsari et al. 2019).

Text classification has variety of applications in the areas like sentiment analysis (Domeniconi et al. 2016; Ramya et al. 2019), document classification (Aubaid and Mishra 2020; Shanavas et al. 2020), spam detection (Pragna and Ramabai 2019; Taninpong and Ngamsuriyaroj 2018), and fake news detection (Kaur, Boparai, and Singh 2019). The automated text categorization using Machine Learning techniques provides more effectiveness, and reduces the efforts needed along with the portability to different fields (Sebastiani 2002). The rapid generation of user contents in large volume in terms of digital information emerges into a need of effective text-classification tactics. The machine learning based methods such as the support vector machines (SVM) (Islam et al. 2020; Kurnia, Tangkuman, and Girsang 2020; H. Liu 2019), naive Bayes(Kavadi, Ravikumar, and Srinivasa Rao 2020; Sueno, Gerardo, and Medina 2020; Tang, Li, and Li 2020; Verma 2020), decision trees, and k-nearest neighbors(Manisha, Kodali, and Srilakshmi 2019; Pragna and Ramabai 2019) are used widely to classify the large amount of text generated online. The false information or misinformation is the inaccurate or incorrect information which could be verified with authentic evidences. It is a challenging task to identify false information and the spreader of false information as it is written with or without the intent to deceive someone (Kumar, West, and Leskovec 2016). The spread of misinformation leads to degradation of quality of information and has wide spread impact on people, business, and all other aspects (Shu

et al. 2018; Zhou and Zafarani 2018; Shariff, Zhang and Sanderso 2017). Misinformation detection is directly related to text classification problem. The machine learning techniques for categorizing the text contents has been significantly achieved but to detect the veracity of the contents has remained as a challenging task. The authors (Ball and Elworthy 2014) extracted features from online repository using part-of-the speech tag and detected misinformation by the classifier logistic regression. The researchers Xu & Guo, explored truthiness of health information related to pro & anti vaccine headlines. The techniques of text mining and sentiment analysis resulted into generation of positive and negative words for pro and anti-vaccine respectively (Xu and Guo 2018).
Deep learning methods have shown significant performance for text classification, however they are expensive for continuously arriving data which might lead to the problem of overfitting as they need to store previous text data in memory and are efficient only on labelled samples(Shan et al. 2020).
However, with the advent of internet the enormous amount of misinformation is generated and the most traditional supervised machine learning methods are not appropriate to handle newly arrived data over the period of time and perform text classification to detect the misinformation. This is because they do not support incremental classification/learning ((Burdisso, Errecalde, and Montes-y-Gómez 2019; Fan and Wang 2012; Wang et al. 2017). An incremental learning deals with large amount of data arriving in the short chunks any time. It improves the performance and accuracy of the knowledge gained as it requires less memory and training time (Chaudhari et al. 2019; Shan et al. 2020). An incremental learning via incremental clustering is ideal for large datasets to achieve best quality clusters(Chaudhari et al. 2019). The proposed incremental clustering algorithm "Distributed Incremental Closeness Factor Based Document Clustering Algorithm" (DICFBADCA) identifies misinformation on the website. This algorithm generates possible Bag-of-Words of synonyms of objectionable words and detects false information using sarcasm detection (Mulay and Joshi 2019).
The analysis of publications can be effectively accomplished using bibliometric analysis. Bibliometric is tool for quantitative and qualitative analysis of research areas, identifying research gaps which helps the potential researchers to analyse the depth of the research done in a field(Chaudhari et al. 2019; Hao et al. 2018). Although, some bibliometric surveys are conducted in the field of text mining in medical research (Hao et al. 2018) and topic modelling (X. Li and Lei 2019), according to the available data it appears that no survey in the study of incremental learning in text classification with the aim of detecting misinformation has been conducted yet. Therefore, the study intends to carry out research in this area by analysing scientific publications from the databases like Scopus, Web of Science, Google Scholar and IEEE. The main objectives are: 1) Publication statistics identification, 2) Collaboration degree analysis 3) Collaboration visualizations 4) Citation Analysis 5) Current techniques applied in this research area.

This bibliometric paper describes the data collection methods, bibliometric statistical analysis, collaboration analysis, publication source analysis, algorithmic analysis, citation analysis, confines of the study, conclusion and future enhancements respectively.

Related Work

Incremental learning in text classification have been applied widely in different area. An incremental learning was applied in the EADS DCS text-mining platform to annotate documents with semantic tags using Conditional Random Fields. The advent of dynamic use of Probabilistic Latent Semantic Analysis with a drift of a factor helps to successfully conduct text classification (Grilheres et al. 2005; L. Zhang et al. 2005).
The SVM based probabilistic classification model was a very powerful technique for text categorization task. An incremental model reduces the training time and is an effective classification model. By using incremental support vector machines it is possible to classify text into multiple classes by updating knowledge of old SVM classifiers, upon arrival of new classes resulting into faster in training the classifier then batch SVM classifiers. Incremental SVM is applied in categorizing large scale corpus of web text. (Cao and Wang 2014; Jia and Mu 2010; B.-F. Zhang, Su, and Xu 2006)
Incremental Perceptron Learning Process is applied to text classification wherein the algorithm ILTC (Incremental Learning of Text Classification) acquires new knowledge by either referring

to old records or forgets previously learned knowledge (Z. Chen, Huang, and Murphey 2007). An evolving Probabilistic Neural Network (ePNN) an incremental Neural Network is used to categorize multi-label text. The ePNN can unceasingly learn without making drastic change to the structure upon arrival of new data(Ciarelli, Oliveira and Salles 2014).

In another study, a cluster based incremental approach was used for spam filtering, forming the clusters of several groups of emails and fetching equal number of keywords from these clusters followed by incremental learning of these clusters upon arrival of new influx of data. Thereby, reducing the cost of retraining the clusters (Hsiao and Chang 2008). An updatable cluster based categorization model was developed using an unsupervised incremental clustering algorithm followed by KNN decision rule to classify text documents(Jiang and Pang 2014). To classify unlabelled text a quick clustering algorithm was developed. A single cluster is formed containing alike unlabelled text-data, then the text set is created by choosing the centre of the text cluster. The incremental learning technique applied further picks ups the text by choosing loss rate between 0 to 1. (Ma, Fan, and Chen 2008).

An incremental learning with fuzzy grammar based approach is another technique to extract structured data from an unstructured text which helps to classify text and match the correct instances. A fuzzy grammar is also applied in supervised learning setting maintaining the coverage of newly arrived data with past data without repeating the training process. AN incremental evolving Fuzzy grammars provides efficient representation that feats syntactic and semantic properties of text. The fuzzy measure used identifies the level of association of the piece of text in a semantic grammar class, similarity between two grammars, and their combination. A novel approach proposed contains background nets an undirected weighted graph that reads and understands the contextual association of terms inside the article using fuzzy set. Further, this approach is enhanced using probabilistic reasoning based on conditional probabilities which offers competitive performance with standard text classification algorithms like SVM, kNN and Naïve Bayes. The incremental learning phase applied using background-net involves enriching the newly arriving terms automatically. (Lo and Ding 2012a, 2012b, 2013; Martin, Shen, and Azvine 2008; Sharef, Martin, and Shen 2009; Sharef and Shen 2010). An effective as well as feasible approach uses the combination of Navie-Bayes along with fuzzy partitioning method to obtain the few samples of labelled training data and then incrementally classify the newly arrived data. (L. Liu and Liang 2011).

Primary Data Collection and Methodology

There are large number of publication databases like Scopus, Google Scholar, ScienceDirect, SCImago, etc. There are open access and paid access publications. Scopus is the largest abstract and citation database of peer-reviewed research literature in the field of science, engineering, technology, medicine, social sciences, arts, and humanities (Chaudhari et al., 2019). The Web of Science (WoS) is another largest high quality publication database. In this study, we have explored Scopus and WoS for statistical analysis. The considerable keywords were identified as mentioned in (Table1) based on the title of the paper.

Considerable keywords

Table1. Considerable keywords from ILTCFD
(Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Keyword search pattern	
Master Keyword (AND)	“incremental learning”
Primary Keyword	"false information" OR "misinformation" OR "text classification" OR "text mining" OR “ text categorization” OR “text processing” OR "detect"
Secondary Keyword (AND NOT)	“imaging” OR “images” OR “audio” OR “video” OR “game”

Basic Explorations of Results

In this study we have explored various databases like Scopus, Web of Science, IEEE, and Google Scholar. The (Table 2) shows the number of research articles obtained from these databases.

Table2. *Number of Research Articles retrieved using Keyword Search*
(Source: Scopus, WoS, IEEE, Google Scholar databases accessed on 28th May 2020 for the period of 2000 to 2020)

Sr. No.	Name of the Database	Publications using Master Keyword and Primary Keywords	Relevant Publications after applying Secondary Keywords	Publications from Journals Only
1	Web Of Science	45	42	37
2	Scopus	248	192	30
3	IEEE	68	48	02
4	Google Scholar	40	15	10
	Total Publications	401	297	79

- In Web of Science, topic field was applied to retrieve documents. Science Citation Indexes Expanded, Social Science Citation Index, and Emerging Sources Citation Indexes were chosen to get quality publications. The WoS core collection resulted into 45 publications using master keyword and primary keywords. After filtering by applying secondary keywords 42 documents mainly related to text were obtained.
- In Scopus database Title-Abstract-Keyword were used as retrieval field. The master keyword and primary keywords mentioned above are applied which results into 248 publications. After removal of irrelevant publications containing secondary keywords, 192 publications were identified.
- In IEEE database, “incremental learning” and “text” resulted into 68 documents using master keyword and primary keywords which after removal of secondary keywords resulted into 48 documents.
- Similarly, Google Scholar showed 40 results using master keyword and primary keywords followed by 15 after using secondary keywords.

Therefore, the total publications in the area of Incremental earning in text classification for False Information Detection were 401 using only primary keywords and 297 using secondary keywords. However, these publications include both conferences and journal papers. The quality of journal publications is much higher than that of conferences, hence we filtered the publications further to get core quality papers from journals resulting into 37, 30, 02, and 10 for WoS, Scopus, IEEE, and Google Scholar databases respectively. Further, after manually removing duplicate publications from all these databases, 62 publications were finally identified from year 2000 to 2019 restricting to only English language. It was observed that there were very less relevant publications related to the topic in the databases like IEEE and Google Scholar. Therefore, we restrict the bibliometric analysis of this area to only WoS and Scopus databases. (Table 3) shows the language trends for incremental learning in text classification for false information detection from Scopus database. The Web of Science database includes 45 publications from English Language.

Table3. *ILTCFD Language Pattern*
(Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Sr. No.	Publication Language	Total Number of Publications
1.	English	184
2.	Chinese	7
3.	Spanish	1
	Total	192

Collaboration Degree Analysis

The collaboration degree is defined as the fraction of sum of the authors, institutes or country of collaborating research publications to the total number of collaborating research publications for a particular year in the specific research area (Hao et al. 2018). Formula to compute collaboration degree is defined in eq1.

$$Cd/y=\sum (Ca/Ci/Cc)/Tcp \qquad \text{-----eq1}$$

Where, Cd/y= Collaboration degree for a particular year, Ca= Collaborating authors for a particular publication, Ci= Collaborating institute for a particular publication, Cc= Collaborating country for a particular publication, Tcp=Total number of collaborating publication for a specific year.

Research Network Analysis (RNA)

Research network analysis is the collection of nodes and edges with the aim of investigating the research publication structure. The nodes represent actors, people, or a thing and edges represent interaction, relations, collaborations or links between the nodes. It involves quantitative assessment on relations between the various actors. Collaboration relations between these nodes can be visualize using RNA. The size of the node depicts the count of authors/institutes/country appearing in the research publication. Bigger the node larger the count is. The edges represent the collaboration between authors/institute/country.

Bibliometric Statistical Analysis

Types of Publications

All the published and unpublished publications were considered for this survey from the year 2000 to 2019. In Scopus database, the researchers have published variety of papers in conferences. It is seen from the results that 53.85% of the papers are published in conferences, whereas 43.41% papers are published in articles whereas there are 88% of the publications as articles in WoS database, followed by 11.09% conference papers (Table 4).

Table4. Types of Publications in ILTCFD (Source: Scopus and WoS databases accessed on 28th May 2020 for the period of 2000 to 2020)

Source type	Number of Publication (Scopus)	Percentage of Publication (Scopus) (%)	Number of Publication (WoS)	Percentage of Publications (WoS)(%)	Source type
Conference Paper	98	53.85	5	11.9	Conference Paper
Article	79	43.41	37	88.09	Article
Conference Review	4	2.2	0	0	Conference Review

Yearly Publishing Trends in ILTCFD

The documents are retrieved from Scopus and WoS database consisting of journal papers, conference papers, review articles, etc. from 2000 to 2020. (Figure 1) shows the yearly publication trends in ILTCFD. It is observed that maximum number of publications are from Scopus database.

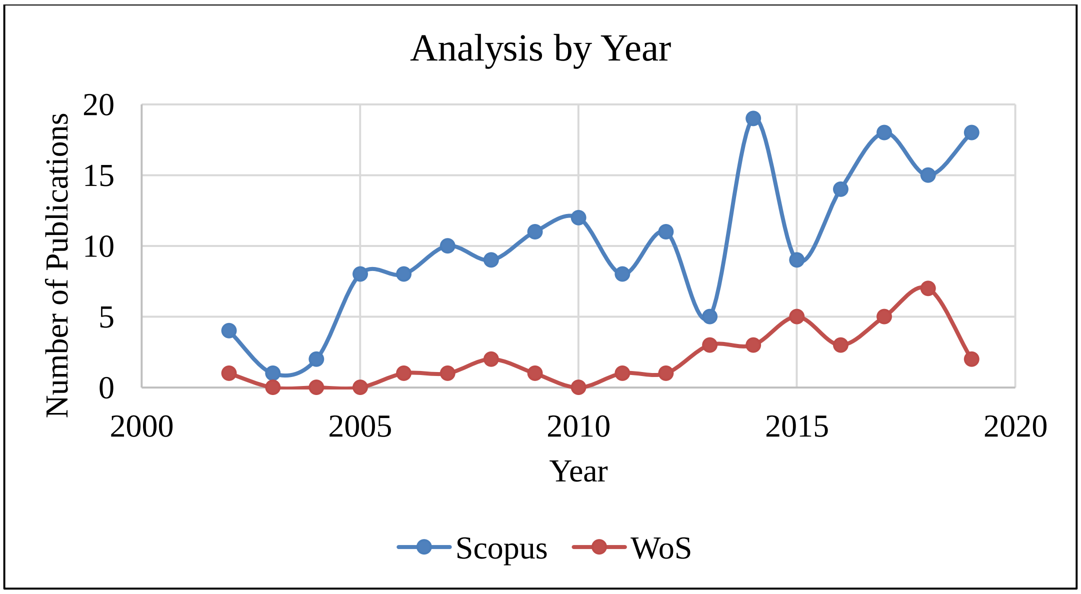

Figure1. Yearly Publication Trends in ILTCFD (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Country Publication

The (Table 5) displays the list of the ten topmost countries producing publications in the field of ILTCFD. It is observed that Chinese lead the publications with nearly 50%. The United States shares 13% of total publications followed by United Kingdom with 11% of publications in both the databases.

Table5. Ten topmost countries producing publications in ILTCFD (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Scopus Database		WoS Database	
Country/ Territory	Number of Publications	Country/ Territory	Number of Publications
China	47	China	12
United States	20	United States	8
United Kingdom	13	United Kingdom	5
Taiwan	11	Brazil	3
Australia	10	India	3
Germany	10	Australia	2
Brazil	10	Germany	2
France	9	Singapore	2
India	8	Spain	2
Japan	8	Taiwan	2

Keyword Statistics

The appropriate keyword does effective searching in the targeted area. The correct combination of keywords is essential for considerable search. The (Table 6) displays first ten keywords list from publications in ILTCFD.

Table6. First Ten Keywords of ILTCFD (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Keywords	Number of Publications
Incremental Learning	141
Learning Systems	59
Learning Algorithms	37
Text Processing	33
Data Mining	31
Classification (of information)	30
Neural Networks	24
Artificial Intelligence	24
Algorithms	19
Support Vector Machines	18

Subject Areas

(Figure 2) shows the various subject areas in which ILTCFD is applied. The major research work of 52% from among all the publications is conducted in the field of Computer Science, 21% of the work is accomplished in the engineering section and 15% of the work is accomplished in the Mathematics section.

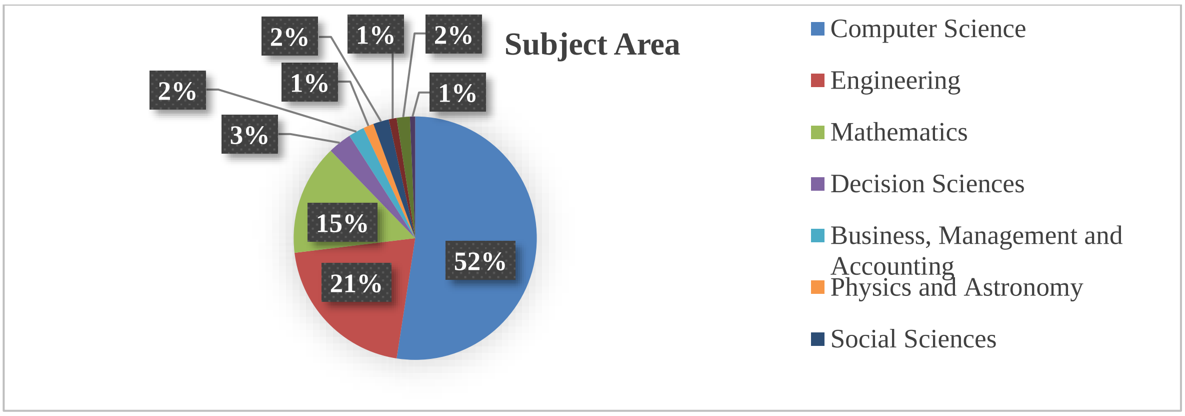

Figure 2. Subject Areas in ILTCFD (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Affiliation Statistics

The top ten universities or organizational affiliations that have contributed in the field of research are displayed in (Figure 3). The Nanjing University, China and National University of Singapore has shown attention towards the field of research in ILTCFD.

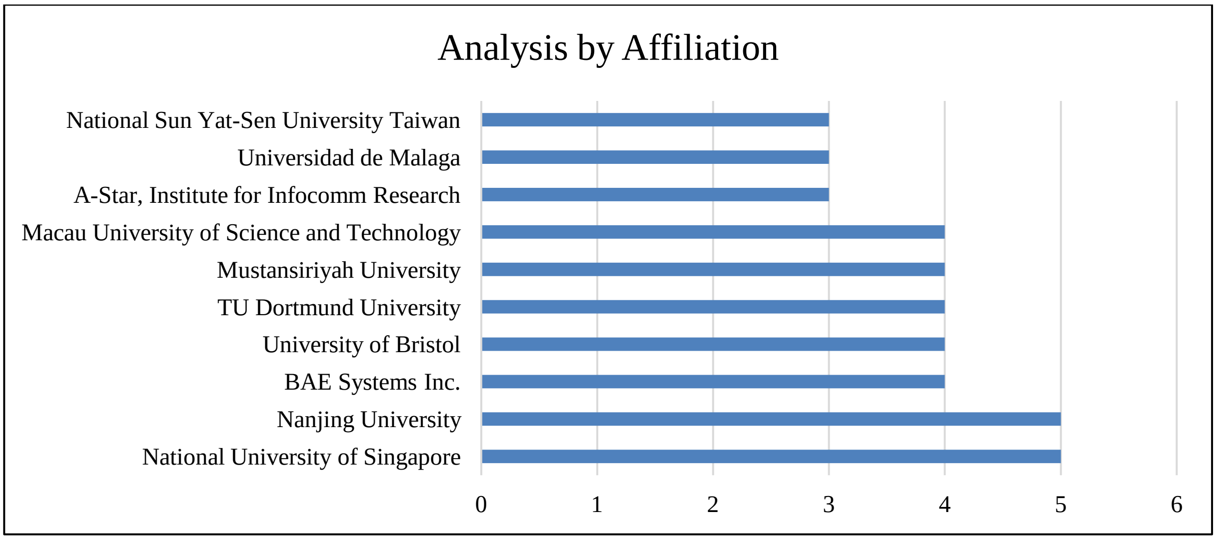

Figure 3. Affiliation Statistics in ILTCFD (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Author Statistics

The bibliometric survey incremental learning in text classification results into 160 authors producing 297 publications. The top ten authors working in the field of research under study and contributing through research publications is presented in (Figure 4). To get the best article in quality journals and its author, we limited the search to articles, and only unpaid access journals for the last 20 years. This results in 79 articles from reputed journals. The top 24 articles were chosen having minimum of 10 citations. The sum of the citations received by the articles based on the number of authors is carried out. (Table 7) displays the citations received, and average number of citations based on the number of authors. It can be observed that maximum citations are received by the articles with two authors, followed by six authors and three authors respectively. Whereas, only 2% of citations are received by four authors articles.

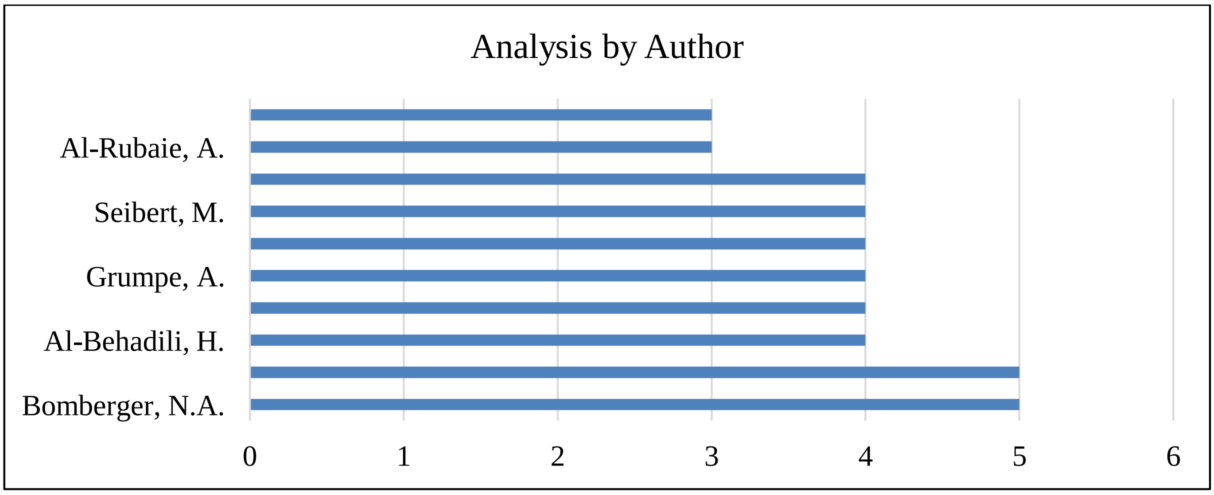

Figure 4: Authors Contribution in ILTCFD (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Table 7. Average Percentage of Citations received based on the number of authors per article in ILTCFD (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Number of Authors	Total Number of Citations	Average Percentage of Citations (Approx.)
1	47	8%
2	151	25%
3	127	21%
4	13	2%
5	101	17%
6	120	20%

Further, the evaluations on the authors with continuous research in the area of incremental learning in text classification results in two productive authors. (Table 8) shows the authors, number of papers published, citations received for the publications, collaborations of the author, and position of author in the article like first author, second author etc., which depicts the position of the author in collaborative publications.

Table8. Productive Authors, citations and collaborative authors in the area of ILTCFD (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Name of the Author	Year	Number of research paper publications	Citations Received	Number of collaborating authors	Position
Zhou, Zhihua	2019	1	1	3	4
	2018	1	9	2	2
	2015	1	22	2	3
	2014	1	106	2	3
	2013	2	38	1	2
Almeida, Tiago	2019	1	0	2	3
	2017	2	19	3,2	3,2
	2016	2	33	3,1	1,1
	2012	2	11	1,1	1,1

Funding Sponsors

(Figure 5) shows the top ten funding sponsors in the area of ILTCFD. It could be observed from the figure that National Natural Foundation of China has provided maximum funding for 14 projects.

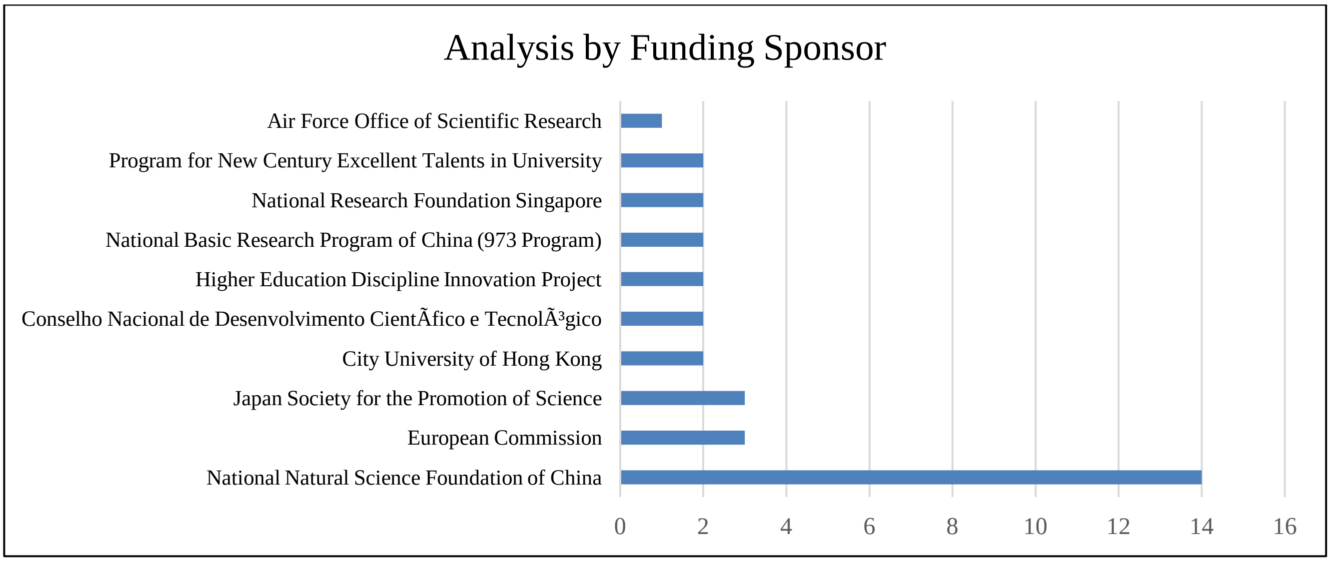

Figure 5. Top Ten Funding Sponsors (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Collaboration Analysis

Author-wise collaboration degree analysis

(Figure 6) presents the annual author-wise collaboration degree analysis. The line chart displays collaboration degree where as bar chart depicts number of collaborative publications per year. Initially the collaboration degree remained steady till 2017, later increases superficially, up to 4.67 in 2018. The average collaboration degree of authors is 3.38.

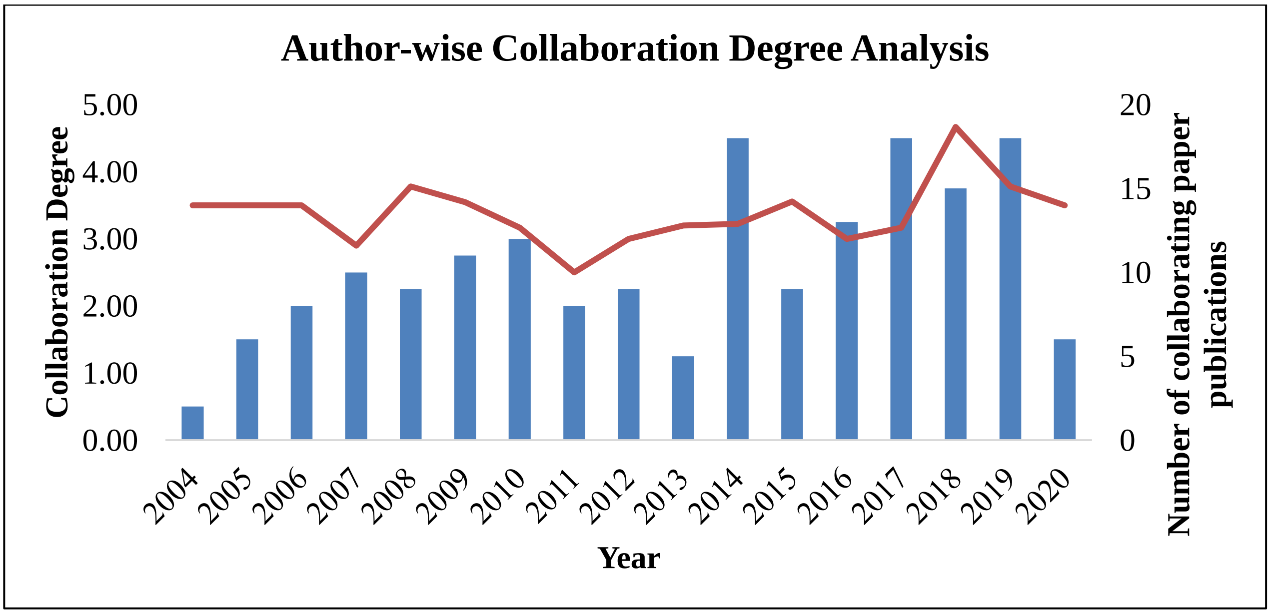

Figure 6. Author-wise collaboration degree analysis (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Country-wise collaboration degree analysis

(Figure 7) depicts the country-wise collaboration degree analysis. The line chart displays collaboration degree where as bar chart depicts number of collaborative publications per year. In 2008 and 2011 collaboration reaches 1 with only 2 and 4 publications respectively. Later, it falls gradually till 2018 with increase in the number of publications. The average country collaboration degree is 0.6. This shows that authors tend to collaborate more within countries compared to international collaborations.

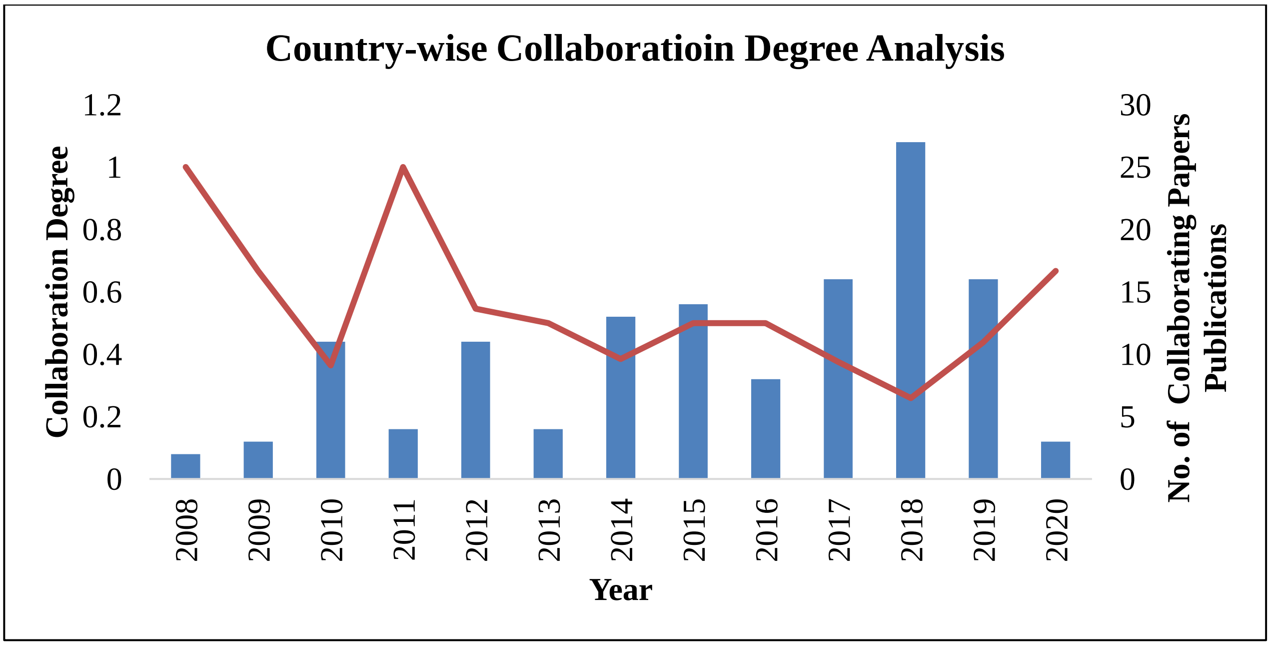

Figure 7. Country-wise collaboration degree analysis (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Country Collaborations Visualizations

(Figure 8) Presents the country-wise collaboration network of 25 nodes representing countries and edges depicting number of collaborations among these countries. The size of the edge among the nodes depict number of collaborations. It is observed that China (the largest node) has the maximum number of collaborations with other countries, followed by USA, Hong Kong and Australia.

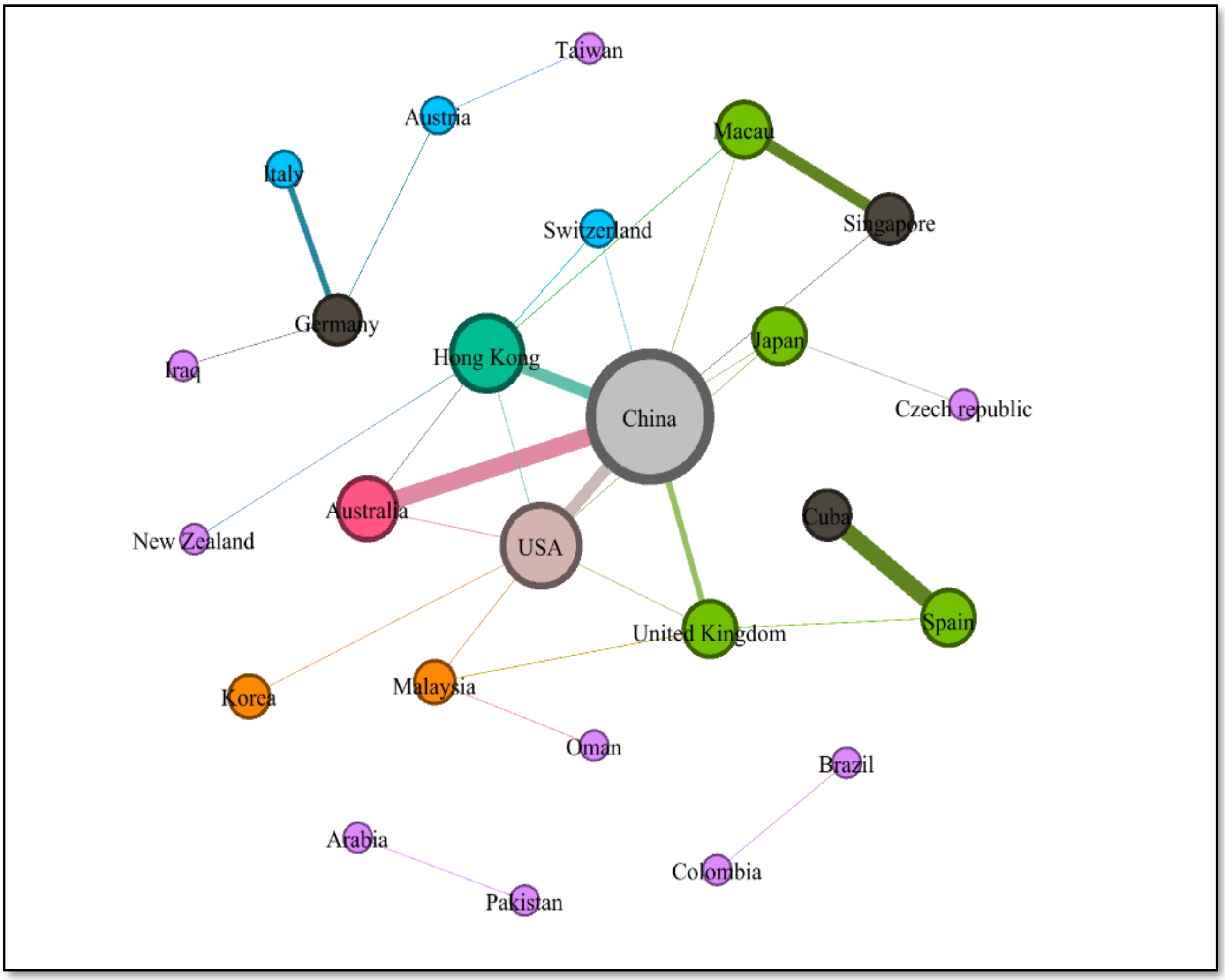

Figure 8. Country-wise collaboration network analysis (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Publication Source Analysis

It is a challenging task to choose a right journal for the publication in the respective research area. Based on the bibliometric analysis done in this paper, the authors recommend following journals which could be used to publish quality research articles. These journals are chosen by restricting the publications to only articles in unpaid journals over the duration of 20 years from 2000 to 2020 which comes to 72 best articles in the area of incremental learning in text classification. Thus, removing conference papers and review, the maximum number of citations received by the journal is computed based on the articles published in that journal and also the reference of such quality articles is provided. (Table 9) shows the top journals with the number and list of quality research papers published in those journals and the total number of citations received by the journal.

Table 9. Top 7 Journals for publications in ILTCFD (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Sr. No.	Name of the Journal	Number of research article	Article Reference	Total number of citations
1	Expert Systems with Applications	5	Hsiao & Chang, 2008, Kho, Lee, Choi, & Kim, 2019, Liu & Ban, 2015, Mena-Torres & & Aguilar-Ruiz, 2014, and Vallim, Andrade Filho, De Mello, & De Carvalho, 2013.	112
2	IEEE Transactions on Knowledge and Data Engineering	3	Frías-Blanco, Del Campo-Ávila, Ramos-Jiménez,Morales-Bueno,Ortiz-Díaz, & Caballero-Mota, 2015, Kasten, & McKinley, 2007, and Zhu, Ting, & Zhou, 2018.	88
3	Pattern Recognition	3	Dewan, Granger, Marcialis, Sabourin, & Roli, 2016, Fan, Zhang, Mei, Peng, & Gao, 2015, and Kang, & Cho, 2009.	73
4	Engineering Applications of Artificial Intelligence	2	Chen, Chang, & Wu, 2016, and Xu, Zhou, & Chen, 2018	41

5	Knowledge-Based Systems	2	Silva, Almeida, & Yamakami, 2017 and Wang, Xu, Lee, & Lee, 2018.	20
6	Journal of Systems and Software	2	Vu, Park, Lee, Lee, Lee, & Ryu, 2010 and Liu, Cukic, Fuller, Yerramalla, & Gururajan, 2006.	17
7	Applied Soft Computing Journal	2	Mohd, & Martin, 2015 and Wang, & Al-Rubaie, 2015.	9

Algorithmic Analysis

(Table 10) shows the broad categories of algorithms and their task or applications in incremental learning in text classification for last 5 years from 2015-2020. In the machine learning category 11 algorithms were identified performing various tasks like text categorization, document classification, topic modelling, proposed algorithm for misinformation detection and spam mail classification and detection.

Table 10. Broad Categories of Incremental Learning algorithms in Text Classification from 2015-2020 (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Sr. No.	Category	Task/Application	Algorithm	Technique	References
1		Misinformation detection	DICFBADCA	Incremental Clutsering, TF-IDF, Context Closness factor, sarcasm detection	(Mulay and Joshi, 2019)
		Text Classification	Learn#	Reinforcement learning, LSTM, ReLU	(Shan et al., 2020)
		Memetic Feature Selection for Multi-Label Text categorization	Novel memetic feature seelction algorithm	Population based incremental learning	(Lee et al., 2019)
		Text Classification	Hybrid MDLText	Minimum Description Length Principle and semantic indexing, short text messages	(Silva, Almeida, et al., 2017)
	Machine Learning	Text Categorization	MDLText	Minimum Description Length Principle and semantic indexing	(Silva, Alberto, et al., 2017)
		Crime text categorization	Evolving Fuzzy grammar	Fuzzy Grammars	(Sharef and Martin, 2015)
		Spam mail classification	Novel online spam classification method	KNN, Naïve Bayes, SVM, Term frequency based interest sets	(Feng et al., 2016)
		Document classification	Ensemble Text Classifer, PCA, KNN, Expectation Maximization	Ensemble Classifier, Unsupervised Learning	(Silambarasan and Anvar Shathik, 2017)
		Topic Modelling for Spam detection	Probabilistic topic modeling	Labeled latent Dirichlet allocation (L-LDA)	(Li et al., 2018)
		Topic Modeling	Partial supervision for HDP	Semi-suoervised learning, Hierarchical Dirichlet Process,	(Wang and Al-Rubaie, 2015)
2	Text Mining	Spam mail classification	Single pass scan algorithm	Tree-based clustering	(Taninpong and Ngamsuriyaroj, 2018)
		Topic Modelling	WIPLSA (Weighted Incremental PLSA)	Probabilistic Latent Semantic Analysis, Weighted Incremental learning	(Li et al., 2018)
		Real time traffic detection in social media	tweet-LDA	Latent Dirichlet Allocation	(Wang et al., 2017)
		Security Market Survelliance	Intelligent agent, Multi-agent based prototype	Rule based reasoning	(Chen et al., 2017)
3	Deep Neural Network	Spam Filter	DBB-RDNN-ReL	N-gram TF-IDF, deep multi layer perceptron	(Barushka and Hajek, 2018)
4	Neural Network	Text Classification	LFNN ——	Back Propagation Lion (BPLion) Neural Network, Fuzzy bounding, Lion algorithm	(Ranjan and Prasad, 2018)
		Document classification	Monarch Butterfly optimization–FireFly optimization based Neural Network	Cluster based indexing	(Kayest and Jain, 2019)

Machine Learning based Incremental Learning algorithms in Text Classification

The parameter free incremental clustering algorithm Distributed Incremental Closeness Factor Based Document Clustering Algorithm (DICFBADCA) is developed to identify misinformation on the website. This algorithm generates possible Bag-of_Words of synonyms of objectionable words and detects false information using sarcasm detection (Mulay and Joshi 2019).

Learn# is an incremental learning algorithm for text classification using reinforcement learning module. It uses techniques of LSTM and Softmax classifier for word-vector creation. In second phase it used Markov decision process and stochastic gradient descent for reinforcement. Due to incremental learning the storage space grows linearly, but becomes stable after using discriminator model. The time required for incremental training is dropped by 80% compared to one-time training (Shan et al. 2020).

The conventional text-mining applications apply feature selection methods for multi-label text categorization. Recently, memetic feature selection methods have become popular due to feature wrappers and filter applied for feature selection. These methods suffer from low performance as it requires problem transformation. The author propose a novel and effective approach for memetic feature selection which outperforms the traditional feature selection methods (Lee et al. 2019). Considering the large amount of data flowing over the web, the MDLText algorithm, a multinomial text classifier, uses minimum description length principle by applying fastest incremental learning techniques and avoids the problem of overfitting. This makes the algorithm robust and more efficient and scalable. ML-MDLText is an extended version of MDLText which involves multi-label text classification using minimum description length principle(Bittencourt, Silva, and Almeida 2019; Silva et al. 2017; Silva, Almeida, and Yamakami 2017).

In another approach the text related to crime is categorized by choosing the text fragments and converted into fuzzy grammars. These evolving-learned grammars are then trained incrementally and adapt to the changes by making small modifications in the model, thus avoiding entire redevelopment of the model. (Mohd Sharef and Martin 2015).

The proposed algorithm on novel spam classification, classifies emails by using techniques like term frequency and Naive Bayes classifier. It uses active learning theory for labelling the emails by users, and finally the incremental learning is applied to retrain the appropriately categorized emails. The algorithm reduces the time of email classification (Feng, Wang, and Zuo 2016).

Another approach, the Ensemble Text Classifier (ETC) is a multistep learning framework for classifying the novel classes from regularized classes in the document classification setting. It uses k-nearest neighbors, infrequent principal component analysis, and expectation maximization which detects the features and prepares the features set (Silambarasan and Anvar Shathik 2017).

Hierarchical Dirichlet Process is an unsupervised learning algorithm widely used for text extraction and text classification. It computes the number of clusters essentially and is flexible in number of topics to be classified. The granular computing has been applied to HDP, which arranges information together according to the similarity or coherence incrementally. Thus resulting in better prediction accuracy (Wang and Al-Rubaie 2015).

A spam filter using text clustering is based on a tree structure where every sub-tree represents one document cluster. As the new document arrives, the incremental learning begins, the newly arrived document is injected into the tree and the clusters are restructured to fit the newly arrived data.(Taninpong and Ngamsuriyaroj 2018).

Text Mining based Incremental Learning algorithms in Text Classification

A dynamic approach to determine the topics and learn incrementally of a topic from the newly arrived document was proposed by a novel Weighted Incremental PLSA algorithm called WIPLSA (N. Li et al. 2018). However, PLSA models suffer from the problem of inferencing new documents. LDA-based approach ('tweet-LDA') was proposed for classification of traffic-related tweets (Wang et al. 2014, 2017). By applying granular computing to Hierarchical Dirichlet process (HDP) for partially supervised data, it could be categorized which adapts to the latest available information. Incremental learning techniques are easy to manage and model the system(Wang and Al-Rubaie 2015). A combination of incremental learning and multiple agent's architecture was proposed for market surveillance systems. The prototype system built includes the techniques of text mining and rule-based reasoning (K. Chen et al. 2017).

Deep Neural Network Based Incremental Learning algorithms in Text Classification

DBB-RDNN-ReL algorithm is a spam filter which overcomes the problem of convergence to poor local minimum and overfitting using incremental learning to handle high dimensional data. It uses n-gram tf-idf feature selection methods along with deep multi-layer perceptron neural network to capture more complex features from high dimensional data, thus no need of high dimensionality reduction to be performed separately (Barushka and Hajek 2018).

LFNN based incremental learning algorithm adopts Back Propagation Lion NN, including fuzzy bounding and Lion algorithm for feasible selection of weights. The algorithm uses context semantic features for text classification (Ranjan and Prasad 2018).

Neural Network Based Incremental Learning algorithms in Text Classification

Monarch Butterfly optimization–FireFly optimization based Neural Network (MB–FF based NN) is another innovative technique used for text classification. The TF-IDF technique is used for extracting features followed by holo-entropy to identify the keywords of the document. Then, MB-FF algorithm is used to index the clusters and finally, documents are retrieved using Bhattacharya distance measure. (Kayest and Jain 2019).

Citation Analysis

(Table 11) shows yearly citations obtained through publications extracted in the area of ILFIDC. The total citation count of 181 publications is 1331 to date. The list of the first ten papers along citations received to them till the date of the data extracted for this research is in (Table 12).

Table 11. Analysis for citations of publications in ILFIDC (Source: Scopus and WoS database accessed on 28th May 2020 for the period of 2000 to 2020)

Year	<2010	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	>2019	total
No. of Citations	83	34	46	49	75	82	112	123	125	194	278	130	1331

Table 12. A citation analysis of top ten publications in ILFIDC (Source: Scopus database accessed on 28th May 2020 for the period of 2000 to 2020)

Title of Publication	Citations received by the publications per Year											
	<2010	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	>2019	Total
MoodLens: An emoticon-based sentiment analysis system for chinese tweets	-	-	-	-	6	19	30	28	23	24	19	4	153
Catch the moment: Maintaining closed frequent itemsets over a data stream sliding window	14	10	13	8	12	13	11	5	9	4	8	1	108
Online and non-parametric drift detection methods based on Hoeffding's bounds	-	-	-	-	-	-	2	5	9	27	26	13	82
Associative learning of vessel motion patterns for maritime situation awareness	14	2	1	4	7	5	3	2	2	7	6	3	56
A similarity-based approach for data stream classification	-	-	-	-	-	4	6	10	11	5	5	1	42
Combining winnow and orthogonal sparse bigrams for incremental spam filtering	10	4	3	7	8	2	2	4	-	1	-	1	42
An incremental cluster-based approach to spam filtering	6	4	4	4	4	2	1	4	3	4	3	1	40
Incremental Learning of Object Detectors without Catastrophic Forgetting	-	-	-	-	-	-	-	-	-	11	18	8	37
A population-based incremental learning approach with artificial immune system for network intrusion detection	-	-	-	-	-	-	-	1	4	9	13	7	34
A class-incremental learning method for multi-class support vector machines in text classification	2	1	3	-	-	6	3	4	3	3	3	3	31

Limitations of the study

This study was conducted from the publication of Scopus and Web of Science databases. The combination of keywords are used for analysis purpose. The research is limited by its query keywords. Hence, there could be few chances that some of the critical articles are missed. However, we believe that these few articles shall not affect our results. The research papers from English language only were considered for study.

Conclusion and Future Enhancements

The study on bibliometric survey on incremental learning in text classification in false information detection and combating results into following findings.

First, the publications are mostly from conferences and journals. These publications are mostly associated with computer science area. The country China has maximum publications in this area. The USA and UK follow China in the publications respectively.

Second, the authors Zhou, Zhihua, and Almedia Tiago having productive work for last 10 years in the area of incremental learning in text classification. The results show the collaborative authors in publications and the total citations received by these authors. The citation analysis by author shows that the citations received by the articles with two authors are large, followed by six authors and three authors respectively.

Third, the seven top most journals namely Expert Systems with Applications, IEEE Transactions on Knowledge and Data Engineering , Pattern Recognition, Engineering Applications of Artificial Intelligence, Knowledge Based Systems, Journal of Systems and Software, and Applied soft computing are identified which could be used for future research and publishing articles.

Fourth, maximum publications are based on techniques from machine learning, followed by deep learning and neural networks. However, there is scope for research in the area of incremental learning for misinformation detection using text classification. The deep learning and neural networks based algorithms using techniques like RNN and CNN could be applied for misinformation detection of shorter text appearing sequentially but they require larger memory space and training time which ultimately affects performance of the algorithm. Hence, the parameter free incremental clustering algorithms like CFBA can be applied for misinformation detection using feature based techniques.

Further the CROWN indicator which gives field normalized citation score and Impact Factor can also be computed.

References

Al-Diabat, M. 2012. "Arabic Text Categorization Using Classification Rule Mining." *Applied Mathematical Sciences* 6(81–84): 4033–46.

Aubaid, A M, and A Mishra. 2020. "A Rule-Based Approach to Embedding Techniques for Text Document Classification." *Applied Sciences (Switzerland)* 10(11).

Ball, L, and J Elworthy. 2014. "Exploring D." *Journal of Marketing Analytics* 2(3): 187–201. doi:10.1057/jma.2014.15

Barushka, Aliaksandr, and Petr Hajek. 2018. "Spam Filtering Using Integrated Distribution-Based Balancing Approach and Regularized Deep Neural Networks." *Applied Intelligence* 48(10): 3538–56. doi:10.1007/s10489-018-1161-y

Bittencourt, M M, R M Silva, and T A Almeida. 2019. "ML-MDLText: A Multilabel Text Categorization Technique with Incremental Learning." In Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019, Institute of Electrical and Electronics Engineers Inc., 580–85. doi:10.1109/BRACIS.2019.00107

Burdisso, S G, M Errecalde, and M Montes-y-Gómez. 2019. "A Text Classification Framework for Simple and Effective Early Depression Detection over Social Media Streams." *Expert Systems with Applications* 133: 182–97.

Cao, J, and H Wang. 2014. "An Improved Incremental Learning Algorithm for Text Categorization Using Support Vector Machine." *Journal of Chemical and Pharmaceutical Research* 6(6): 210–17.

Chaudhari, A, R R Joshi, P Mulay, K Kotecha, and P Kulkarni. 2019. Bibliometric survey on incremental clustering algorithms. *Library Philosophy and Practice*, 2019.

Chen, K, X Li, B Xu, J Yan, and H Wang. 2017. Intelligent agents for adaptive security market surveillance. *Enterprise Information Systems*, 11(5): 652–671. doi: 10.1080/17517575.2015.1075593

Chen, Z, L Huang, and Y L Murphey. 2007. "Incremental Learning for Text Document Classification." In IEEE International Conference on Neural Networks - Conference Proceedings, Orlando, FL, 2592–97. doi:10.1109/IJCNN.2007.4371367

Ciarelli, P M, E Oliveira, and E O T Salles. 2014. "Multi-Label Incremental Learning Applied to Web Page Categorization." *Neural Computing and Applications* 24(6): 1403–19.doi: 10.1007/s00521-013-1345-7

Domeniconi, G, G Moro, R Pasolini, and C Sartori. 2016. "A Comparison of Term Weighting Schemes for Text Classification and Sentiment Analysis with a Supervised Variant of Tf.Idf" ed. Helfert M Francalanci C Belo O. Holzinger A. *Communications in Computer and Information Science* 584: 39–58.

Fan, Xinghua, and Shaozhu Wang. 2012. "An Incremental Learning Algorithm Considering Texts' Reliability." *International Journal of Advanced Computer Science and Applications* 3(2): 23–28.

Feng, Lizhou, Youwei Wang, and Wanli Zuo. 2016. "Quick Online Spam Classification Method Based on Active and Incremental Learning." *Journal of Intelligent and Fuzzy Systems* 30(1): 17–27. doi:10.3233/IFS-151707

Grilheres, B, S Canu, C Beauce, and S Brunessaux. 2005. "A Platform for Semantic Annotations and Ontology Population Using Conditional Random Fields." In Proceedings - 2005 IEEE/WIC/ACM InternationalConference on Web Intelligence, WI 2005, Compiegne Cedex, 790–93. doi:10.1109/WI.2005.10

Guo, Y, Z Shao, and N Hua. 2010. "Automatic Text Categorization Based on Content Analysis with Cognitive Situation Models." *Information Sciences* 180(5): 613–30. doi:10.1016/j.ins.2009.11.012

Hao, Tianyong, Xieling Chen, Guozheng Li, and Jun Yan. 2018. "A Bibliometric Analysis of Text Mining in Medical Research." *Soft Computing* 22(23, SI): 7875–92. . doi:10.1007/s00500-018-3511-4

Hsiao, W.-F., and T.-M. Chang. 2008. "An Incremental Cluster-Based Approach to Spam Filtering." *Expert Systems with Applications* 34(3): 1599–1608. doi:10.1016/j.eswa.2007.01.018

Islam, M K, M M Ahmed, K Z Zamli, and S Mehbub. 2020. "An Online Framework for Civil Unrest Prediction Using Tweet Stream Based on Tweet Weight and Event Diffusion." *Journal of Information and Communication Technology* 19(1): 65–101.

Jia, Z, and J Mu. 2010. "Web Text Categorization for Large-Scale Corpus." In ICCASM 2010 - 2010 International Conference on Computer Application and System Modeling, Proceedings, Shanxi, Taiyuan, V8188–91. doi:10.1109/ICCASM.2010.5619341

Jiang, S, and G Pang. 2014. "Text Categorization Based on Unsupervised Incremental Clustering and KNN." *Journal of Computational Information Systems* 10(1): 171–79. doi: 10.12733/jcis8672

Kaur, P, R S Boparai, and D Singh. 2019. "Hybrid Text Classification Method for Fake News Detection."

International Journal of Engineering and Advanced Technology 8(5): 2388–92.

Kavadi, D P, P Ravikumar, and K Srinivasa Rao. 2020. "A New Supervised Term Weight Measure for Text Classification." *International Journal of Advanced Science and Technology* 29(6): 3115–28.

Kayest, M, and S K Jain. 2019. "An Incremental Learning Approach for the Text Categorization Using Hybrid Optimization." *International Journal of Intelligent Computing and Cybernetics* 12(3): 333–51. doi:10.1108/IJICC-12-2018-0170

Kowsari, K, K J Meimandi, M Heidarysafa, S Mendu, L Barnes, and D Brown. 2019. Text classification algorithms: A survey. *Information (Switzerland)*, 10(4). doi:10.3390/info10040150

Kurnia, R I, Y D Tangkuman, and A S Girsang. 2020. "Classification of User Comment Using Word2vec and SVM Classifier." *International Journal of Advanced Trends in Computer Science and Engineering* 9(1): 643–48.

Lakshmi Prasanna, P, and D Rajeswara Rao. 2017. "Literature Survey on Text Classification: A Review." *Journal of Advanced Research in Dynamical and Control Systems* 9(Special Issue 12): 2270–80.

Lee, J, I Yu, J Park, and D.-W. Kim. 2019. "Memetic Feature Selection for Multilabel Text Categorization Using Label Frequency Difference." *Information Sciences* 485: 263–80. doi:10.1016/j.ins.2019.02.021

Li, N, W Luo, K Yang, F Zhuang, Q He, and Z Shi. 2018. Self-organizing weighted incremental probabilistic latent semantic analysis. *International Journal of Machine Learning and Cybernetics*, 9(12): 1987–1998. doi: 10.1007/s13042-017-0681-9

Li, X, and L Lei. 2019. "A Bibliometric Analysis of Topic Modelling Studies (2000–2017)." *Journal of Information Science.*

Liu, H. 2019. "A Location Independent Machine Learning Approach for Early Fake News Detection." In Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, ed. Khan L Hu X T Ak R Tian Y Barga R Zaniolo C Lee K Ye Y F Baru C. Huan J. Institute of Electrical and Electronics Engineers Inc., 4740–46.

Liu, L, and Q Liang. 2011. "A High-Performing Comprehensive Learning Algorithm for Text Classification without Pre-Labeled Training Set." *Knowledge and Information Systems* 29(3): 727–38. doi: 10.1007/s10115-011-0387-3

Lo, S.-L., and L Ding. 2012a. "Learning and Reasoning on Background Net - Is Application to Text Categorization." ICIC Express Letters 6(3): 625–31.

Lo, S.-L., and L Ding. 2013. "Learning and Reasoning on Background Nets for Text Categorization with Changing Domain and Personalized Criteria." *International Journal of Innovative Computing, Information and Control*, 9(1), 47–67.

Lo, S.-L., and L Ding. 2012b. 2012b. "Probabilistic Reasoning on Background Net: An Application to Text Categorization." In Proceedings - International Conference on Machine Learning and Cybernetics, Xian, Shaanxi, 688–94. doi:10.1109/ICMLC.2012.6359008

Ma, H, X Fan, and J Chen. 2008. "An Incremental Chinese Text Classification Algorithm Based on Quick Clustering." In Proceedings - International Symposium on Information Processing, ISIP 2008 and International Pacific Workshop on Web Mining and Web-Based Application, WMWA 2008, Moscow, 308–12. doi: 10.1109/ISIP.2008.126

Manisha, M, A Kodali, and V Srilakshmi. 2019. "Machine Classification for Suicide Ideation Detection on Twitter." *International Journal of Innovative Technology and Exploring Engineering* 8(12): 4154–60.

Martin, T, Y Shen, and B Azvine. 2008. "Incremental Evolution of Fuzzy Grammar Fragments to Enhance Instance Matching and Text Mining." *IEEE Transactions on Fuzzy Systems* 16(6): 1425–38. doi: 10.1109/TFUZZ.2008.925920

Mohd Sharef, N, and T Martin. 2015. "Evolving Fuzzy Grammar for Crime Texts Categorization." *Applied Soft Computing Journal* 28: 175–87. doi:10.1016/j.asoc.2014.11.038

Mulay, P, and R R Joshi. 2019. "Journey of CFBA Variants with Advancement in Text-Mining and Subspace-Clustering." *International Journal of Scientific and Technology Research* 8(8): 467–73.

Pragna, B, and M Ramabai. 2019. "Spam Detection Using NLP Techniques." *International Journal of Recent Technology and Engineering* 8(2 Special Issue 11): 2423–26.

Ramya, P, G Guttikonda, V Sindhura, and V K Gadde. 2019. "Performance Based Machine Learning Algorithm for Topic Oriented Text Categorization." *International Journal of Recent Technology and Engineering* 8(2 Special Issue 11): 3501–6.

Ranjan, N M, and R S Prasad. 2018. "LFNN: Lion Fuzzy Neural Network-Based Evolutionary Model for Text Classification Using Context and Sense Based Features." *Applied Soft Computing Journal* 71: 994–1008. doi:10.1016/j.asoc.2018.07.016

Sebastiani, F. 2002. "Machine Learning in Automated Text Categorization". *ACM Computing Surveys*, 34(1): 1–47. doi:10.1145/505282.505283

Shan, G, S Xu, L Yang, S Jia, and Y Xiang. 2020. Learn#: A Novel incremental learning method for text classification. *Expert Systems with Applications*, 147. doi: 10.1016/j.eswa.2020.113198

Shanavas, N, H Wang, Z Lin, and G Hawe. 2020. "Ontology-Based Enriched Concept Graphs for Medical Document Classification." *Information Sciences* 525: 172–81.

Sharef, N M, T Martin, and Y Shen. 2009. "Minimal Combination for Incremental Grammar Fragment Learning." In 2009 International Fuzzy Systems Association World Congress and 2009 European Society for Fuzzy Logic and Technology Conference, IFSA-EUSFLAT 2009 - Proceedings, Lisbon, 909–14.

Sharef, N M, and Y Shen. 2010. "Text Fragment Extraction Using Incremental Evolving Fuzzy Grammar Fragments Learner." In 2010 IEEE World Congress on Computational Intelligence, WCCI 2010, Barcelona. doi:

10.1109/FUZZY.2010.5584010

Sharef, N M, and T Martin. 2015. Evolving fuzzy grammar for crime texts categorization. *Applied Soft Computing*, 28: 175–187. doi:10.1016/j.asoc.2014.11.038

Shariff, SM, X Zhang, M Sanderso. 2017. On the Credibility Perception of News on Twitter: Readers, Topics and Features, Computers in Human Behaviour, 75: 785-796.

Shu K, D Mahudeswaran, S Wang, D Lee, H Liu, 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media, *arXiv:1809.01286v1*

Silva, R M, T C Alberto, T A Almeida, and A Yamakami. 2017. "Towards Filtering Undesired Short Text Messages Using an Online Learning Approach with Semantic Indexing." *Expert Systems with Applications 83: 314–25.* doi:10.1016/j.eswa.2017.04.055

Silva, R. M., Almeida, T. A., & Yamakami, A. 2017. MDLText: An efficient and lightweight text classifier. *Knowledge-Based Systems*, 118, 152–164. doi: 10.1016/j.knosys.2016.11.018

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In Proceedings of the 25th International Conference on World Wide Web, WWW '16, pages 591–602, Montreal, Qu ´ ebec, Canada, 2016. International World Wide Web Conferences ´ Steering Committee.

Sueno, H T, B D Gerardo, and R P Medina. 2020. "Multi-Class Document Classification Using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique." *International Journal of Advanced Trends in Computer Science and Engineering* 9(3): 3937–44.

Tang, Z, W Li, and Y Li. 2020. "An Improved Term Weighting Scheme for Text Classification." *Concurrency Computation* 32(9).

Taninpong, P, and S Ngamsuriyaroj. 2018. "Tree-Based Text Stream Clustering with Application to Spam Mail Classification." *International Journal of Data Mining, Modelling and Management* 10(4): 353–70.doi:10.1504/IJDMMM.2018.095354

Verma, P K. 2020. "Exploration of Text Classification Approach to Classify News Classification." *International Journal of Advanced Science and Technology* 29(5): 2555–62.

Wang, D, and A Al-Rubaie. 2015. "Incremental Learning with Partial-Supervision Based on Hierarchical Dirichlet Process and the Application for Document Classification." *Applied Soft Computing Journal* 33: 250–62. doi:10.1016/j.asoc.2015.04.044

Wang, D, A Al-Rubaie, S S Clarke, and J Davies. 2017. "Real-Time Traffic Event Detection from Social Media." *ACM Transactions on Internet Technology* 18(1). doi:10.1145/3122982

Wang, D, A Al-Rubaie, J Davies, and S S Clarke. 2014. "Real Time Road Traffic Monitoring Alert Based on Incremental Learning from Tweets." In IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - EALS 2014: 2014 IEEE Symposium on Evolving and Autonomous Learning Systems, Proceedings, Institute of Electrical and Electronics Engineers Inc., 50–57. doi:10.1109/EALS.2014.7009503

Xu, Z, and H Guo. 2018. "Using Text Mining to Compare Online Pro- and Anti-Vaccine Headlines: Word Usage, Sentiments, and Online Popularity." *Communication Studies* 69(1): 103–22. doi:10.1080/10510974.2017.1414068

Zhang, B.-F., J.-S. Su, and X Xu. 2006. "A Class-Incremental Learning Method for Multi-Class Support Vector Machines in Text Classification." In Proceedings of the 2006 International Conference on Machine Learning and Cybernetics, Dalian, 2581–85. doi:10.1109/ICMLC.2006.258853

Zhang, L, C Li, Y Xu, and B Shi. 2005. "An Efficient Solution to Factor Drifting Problem in the PLSA Model." In Proceedings - Fifth International Conference on Computer and Information Technology, CIT 2005, Shanghai, 175–79. doi:10.1109/CIT.2005.70

Zhou, X., Zafarani, R. 2018. Fake News: A Survey of Research, Detection Methods, and Opportunities, *arXiv:1812.00315v1 [cs.CL]*