Scientific Paper Title Validity Checker Utilizing Vector Space Model and Topics Model

Jan Wira Gotama Putra
Informatics/Computer Science Undergraduate Program
Institut Teknologi Bandung
Bandung, Indonesia
wiragotama@gmail.com

Katsuhide Fujita, Ph.D.
Department of Computer and Information Sciences,
Tokyo University of Agriculture and Technology
Tokyo, Japan
katfuji@cc.tuat.ac.jp

Abstract—**Many efforts has been made to discover ways to understand topics that lies in texts. Especially, scientific papers are one of important targets of understanding topics by analyzing texts because they contain many technical terms and follow the academic writing. In this paper, we apply text analysis methods that includes topics modelling to build a system that could check whether scientific paper title suits its abstract. We utilized two term weighting methods (TF-IDF and BM25), and terms-topics probability model by utilizing *Latent Semantic Indexing* (LSI) and *Latent Dirichlet Allocation* (LDA). We evaluated the models in 3 different domains of dataset. We found out that our model performed quite well despite of some drawbacks. We conclude that our method in title checking could provide robust and consistent performance.**

Keywords— title checker, text analysis.

I. INTRODUCTION

The effort of making artificial intelligence that could understand natural language has been growing for the past decades. Despite of research breakthrough, there are still no research that could make computer understands natural language completely. One of the problems is the inability of computer to understand topic and context of the natural language completely. This aspect actually plays an important role in natural language understanding [1-3].

In artificial intelligence field, topic understanding plays role in many areas such as documents summarization, natural language generation and even information retrieval [4]. Development of topic understanding could lead into further research that enhances the quality of natural language understanding. With no doubt, it is essential in enhancing the quality of future technology [5,6].

Especially, scientific papers are one of important targets of topics understanding by analyzing texts because they contain many technical terms and follow the academic writing. Most people usually put most of their effort in writing the content of paper and only allocate short time to write the title, which makes it less in quality. Meanwhile, title plays an important part to the paper and correlate to the number of downloads and citations [7]. Therefore, it is really important to provide a way to judge whether title really represents the scientific paper to ensure the quality of the title. It would especially help novice writer in writing to ensure that their title depicts its content as title is very important.

In this paper, we propose a work of topic understanding research that implemented as scientific paper title validity checker to check whether scientific paper title matches its abstract. This research also explores more about the potential usage of vector space model and topics model, particularly in title checking task. We use abstract text as it represents the scientific paper content in short length. By analyzing the texts and abstracts of the scientific papers, the computers can check the validity of the scientific paper title, automatically. However, the system utilizing term weighting methods (e.g. TF-IDF and BM25 and so on), and terms-topic probability model (e.g. LSI and LDA and so on) has not evaluated under the large sized datasets. The results of this study play an important role in natural language understanding to the scientific papers, especially in large sized datasets. The checker would be useful in checking scientific paper in several domains as it could minimize human efforts in judging the relatedness between the scientific paper's title and its abstract. The work is evaluated under the scientific paper written in English in three research domains of biochemistry: Gallium Nitride related (GaN), Complex Network (ComNet), and Carbon.

The remainder of the paper is organized as follows. First, we propose the scientific paper title validity checker. Next, we present our experimental analysis. Finally, we present our conclusions.

II. METHODOLOGY

In making scientific paper title, researcher usually place key terms that occurs in abstract, supposedly its topic also matches with it's abstract topic. Taking this heuristic as foundation, we decided to utilize vector space model and topics model. Vector space model provides key terms selection of a text [8,9], meanwhile topics model provides topics-mixture of a text alongside with the probability of each term contribution in topic that could be done through statistical analysis, which based on terms weighting [10, 11].

Checking title to its abstract consisted of several steps. It should be noticed that our document has two parts, title and abstract. The first step is to process the text documents into easier representation for further process, in this particular research; we used words-based representation. The text processing includes splitting, tokenizing, part of speech tagging, lemmatization and stop-words removal. The first step will result in *vector of terms*. We only consider nouns and verbs for further processing. The second step is to make terms-documents weighting matrix to weight each terms based on its occurrence in documents. We utilized *Term-Frequency and Inverse Document Frequency* (TF-IDF) algorithm; and BM25 algorithm for this case [8,9] to make vector space model. The third step is to utilize vector space model generated from TF-IDF algorithm to make terms-topic probability model by utilizing *Latent Semantic Indexing* (LSI) and *Latent Dirichlet Allocation* (LDA) algorithm [10,11]. The next step is utilizing terms-documents weighting matrix and terms-topics probability model in comparing concept signature between title and abstract text which is essential in title-abstract checking. The checking is done in binary classification manner to classify the instance as positive or negative. The steps could be seen in Figure 1.

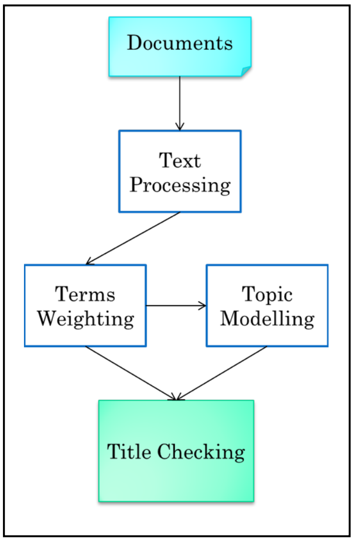


Figure 1. General Workflow of the Methodology

In this study, we propose two judgment methods for checking whether title matches its abstract or not.

A. *Terms Occurrence Based Judgment*

The first judgment is *terms occurrence based judgment*. This analysis involves utilizing vector space model produced by TF-IDF or BM25. Each term occurred in title and abstract will be looked up for its weight in vector space model, producing terms weight vector for title and abstract part of document. All elements of the vector then summed to produce average weight for title and abstract, and then the average weight of the title divided by the average weight of the abstract resulting *terms matching score*. If the *terms matching score* is more than defined *threshold*, then the title would be considered as match to its abstract. This judgment will provide analysis whether key terms in abstract also usually appears in title; or how similar the terms occurred in title to the terms occurred in abstract. This judgment was based on the human behavior that place key terms both in abstract and title.

B. *Topics Based Judgment*

The second judgment method is topics based judgment. This judgment was based on the *heuristic* that supposedly; scientific paper title's topic should match to its abstract. This analysis involves utilizing terms-topics matrix produced by LSI or LDA. In LSI and LDA need vector space model as input. We provide TF-IDF based vector space model as the input.

In LSI, the vector space model is decomposed into 3 matrices U, $\sum$ and V^T. U is terms-topic weight representation matrix, V^T is topic-documents weight representation matrix and $\sum$ is representation of importance in *semantic* dimensions [4]. In this research, U and V^T are reduced and utilized for further step.

In Latent Dirichlet Allocation, topics are associated with terms [11]. It is assumed that one document has various composition of topics, therefore each term in document is part of topics in probability manner. This algorithm explicitly models terms distribution across various topics which assumed to be independent to each other. LDA has capability to determine mixture of topics in a document. LDA will result in terms-terms probability clusters. The clusters are used to construct terms-topics weight matrix. Illustration of terms-topics matrix could be seen in Figure 2.

Each terms occurred in title and abstract will be looked up for its weight in terms-topics weight matrix, producing topics probability vector for title and abstract part of document. The probability score for each topic in the topics probability vector is the average score of probability score of each term that occurred in the respective text (title or abstract). To judge whether title match for the abstract, the *cosine distance* between two topics probability vector will be computed. If the *cosine distance* between two vectors more than defined *threshold,* then the title would be considered as match to its abstract.

	Topic0	Topic1	Topic2	Topic3
Term1	value	value	value	...
Term2	...			
Term3	...			
Term4	...			

Figure 2. Terms-Topics Matrix Illustration

III. EXPERIMENTAL RESULTS

A. *Experimental Setting*

The model was constructed and tested using 3 domains of dataset, normal and large sized. The datasets took from biochemistry research fields: Gallium Nitride (GaN), Complex Network (ComNet) and Carbon. Details about the dataset size could be seen in Table 1.

Table 1. Dataset Details

Normal Sized Dataset			
	GaN	**ComNet**	**Carbon**
Correct Documents	1044	995	950
Wrong Documents	1044	968	1108
Number of distinct terms in documents	4575	13945	13276
LDA Sampling iterations	Number of terms / 2		
Large Sized Dataset			
	GaN	**ComNet**	**Carbon**
Correct Documents	1878	1986	2290

Wrong Documents	1878	2032	2139
Number of distinct terms in documents	5510	20667	17848
LDA sampling iterations	2100		

There are four evaluation metrics that considered into our account: *precision, recall, F-measure* and *accuracy*:

- *Precision= TP/(TP+FP)*
- *Recall=TP/(TP+FN)*
- *F-Measure=2 (Precision x Recall)/(Precision+Recall)*
- *Accuracy= (TP+TN)/(TP+TN+FP+FN)*

(True Positives(TP), False Positives(FP), True Negatives(TN), and *False Negatives(FN))*

We tested the model performance over all dataset domains across the *threshold*, ranged from *0.1-0.9*. In particular, we also tested LSI and LDA based models performance over the number of topics, from *10%* to *100%* to the number of terms.

B. Dataset Corpus

The following are examples of preprocessed positive and negative labeled instances in training corpus (only nouns and verbs). Label=1 means positive instance while label=0 means negative instance.

Author = LAMPL, Y; ESHEL, Y; BENDAVID, E; GILAD, R; SAROVAPINHAS, I; SANDBANK,
Title = [neuropathy, central, nervous, system, manifestation]
Abstract Text = [author, describe, woman, neuropathy, gan, cn, involvement, admit, hospital, generalize, seizure, gait, disturbance, follow, deterioration, childhood, examination, reveal, retardation, scanning, speech, cerebellar, dysfunction, pyramidal, sign, extremity, neuropathy, nerve, conduction, velocity, decrease, brain, ct, mri, show, demyelination, nerve, biopsy, reveal, sign, gan, patient, sister, die, age, disturbance, childhood, case, illustrate, presentation, gan, characterize, neuropathy, cn, involvement, include, seizure]
Label = 1

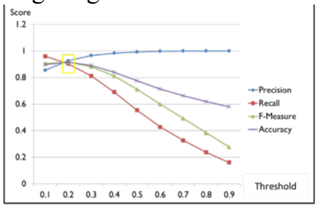
Author = NAKAMURA,
Title = [analysis, monitoring, using, interference, effect]
Abstract Text = [gan, film, obtain, annealing, irradiation, leebus, treatment, show, resistivity, omega, cm, annealing, temperature, 600, degrees, c, case, annealing, temperature, room, temperature, 1000, degrees, c, gan, film, show, change, resistivity, 2, omega, cm, 8, omega, cm, result, indicate, hydrogen, produce, nh3, dissociation, temperature, 400, degrees, c, relate, hole, compensation, mechanism, hydrogenation, process, acceptor, h, complex, form, gan, film, propose, formation, acceptor, h, complex, cause, hole, compensation, emission, photoluminescence]
Label = 0

C. Experimental Result

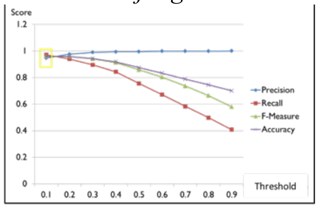
We show the result of the normal and large sized dataset in this paper. The yellow mark on the graph means the best evaluation metrics combination location.

1) TF-IDF and BM-25

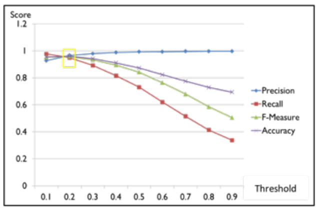
From the figures, it could be seen that the trend for all dataset domains, both normal and large sized, is same for both TF-IDF and BM25 weighting scheme based on frequencies of terms (Figure 3 – 6). We found out the trend for all dataset domains, both normal and large sized, is same for both TF-IDF and BM25 weighting scheme on *terms based occurrence judgment*.



GaN

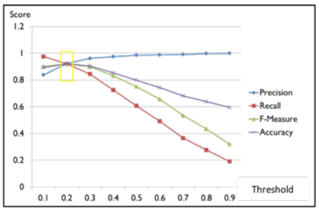


ComNet

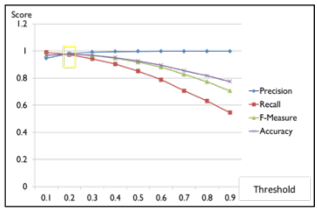


Carbon

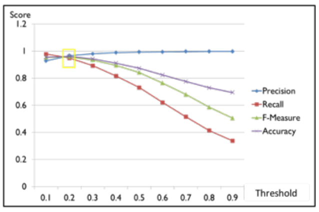
Figure 3. TF-IDF Model Testing Result, Normal Dataset



GaN

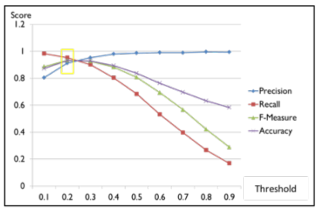


ComNet

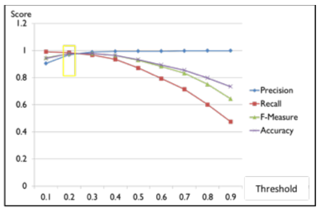


Carbon

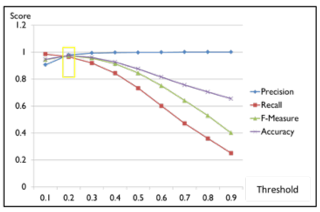
Figure 4. TF-IDF Model Testing Result, Large Dataset



GaN

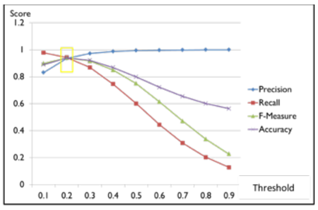


ComNet

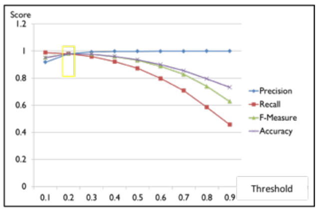


Carbon

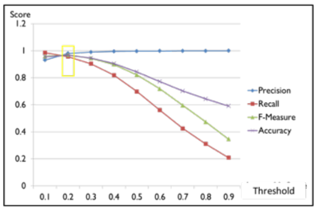
Figure 5. BM25 Model Testing Result, Normal Dataset



GaN



ComNet



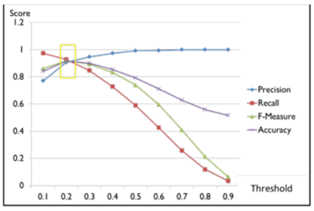
Carbon

Figure 6. BM25 Model Testing Result, Large Dataset

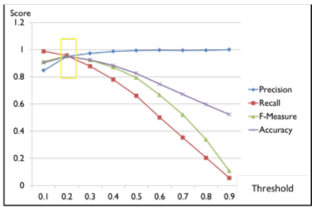
2) LSI and LDA

We found out the best LSI model performance for GaN, ComNet and Carbon when the topics are 20%, 10% and 10% to the number of terms respectively. As for large sized dataset, GaN, ComNet and Carbon performance best when the number of topics are 10% to the number of terms for all dataset. Details could be seen in Figure 7 and Figure 8.

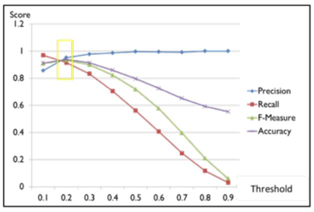
As for LDA, we found out the best model when the number of topics are 20%, 10% and 10% for GaN, ComNet and Carbon large dataset respectively. While the it is 10% for all normal datasets. Details could be seen in Figure 9 and Figure 10.



GaN, Topics = 20% of terms

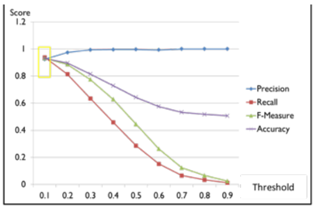


ComNet, Topics = 10% of terms

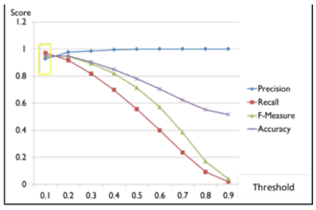


Carbon, Topics = 10% of Terms

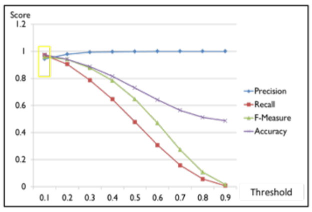
Figure 7. LSI Model Testing Result, Normal Dataset



GaN, Topics = 10% of terms

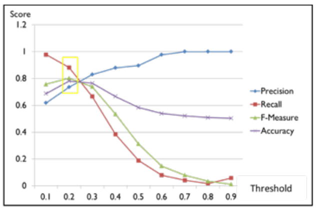


ComNet, Topics = 10% of terms

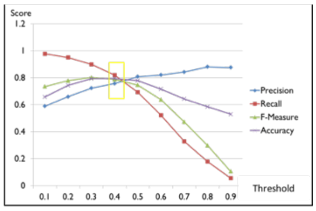


Carbon, Topics = 10% of Terms

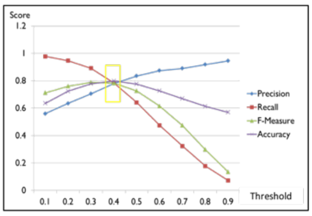
Figure 8. LSI Model Testing Result, Large Dataset



GaN, Topics = 10% of terms

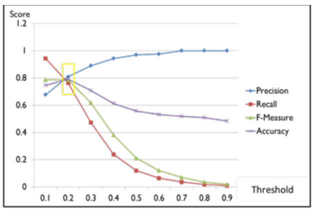


ComNet, Topics = 10% of terms

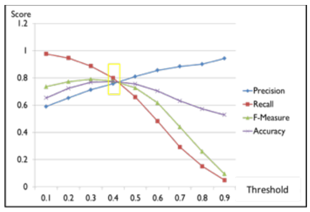


Carbon, Topics = 10% of Terms

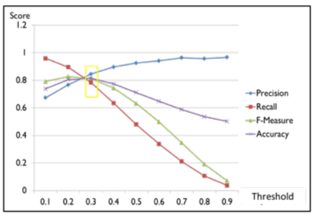
Figure 9. LDA Model Testing Result, Normal Dataset



GaN, Topics = 20% of terms



ComNet, Topics = 10% of terms



Carbon, Topics = 10% of Terms

Figure 10. LDA Model Testing Result, Large Dataset

3) *Discussion*

It could be seen that the trend for all dataset domains, both normal and large sized, is same for both TF-IDF and BM25 weighting scheme based occurrence judgment. When the *threshold* is increased, the evaluation parameters also have bigger gap to each other, means the performance of model decreased. From the testing result, we found that the *terms matching score* is concentrated mostly from *0.1* to *0.6* for correct documents for all algorithms. On the other hand, *terms matching score* is concentrated mostly from *0.0* to *0.1* for wrong documents. It means, key terms that have high weight usually occur both in title and abstract part of the document. As the length of the title usually small compared to the length of the abstract, *terms matching score* that concentrated in 0.1 to 0.6 is reasonable. As the *threshold* increased, less relevant items are correctly classified as most of the instances classified as *negative*. On the other hand, *true positives* and *false positives* instances become fewer, making precision higher while recall become lower. In this case, *accuracy* and *F-measure* play very important role in judging models performance.

On the other hand, LSI and LDA performance usually become worse over increasing number of topics to terms for all dataset domains. In LSI, the model performance decreased as the *threshold* increased. On the other hand, the model performance of LDA rose as the *threshold* increased to a certain value, then it become worse. From the testing result, we found that the *cosine distance score* is concentrated mostly from *0.1* to *0.6* for correct documents for both algorithms. On the other hand, *cosine distance score* is concentrated mostly from *0.0* to *0.1* for correct documents. It means, title and abstract has similar composition of topics distribution probability. As the length of the title usually small compared to the length of the abstract, *cosine distance score* that concentrated in *0.1* to *0.6* is reasonable. As the *threshold* increased, less relevant items are correctly classified as most of the instances classified as *negative*.

We found out the trend for the normal dataset and large dataset is quite similar. Therefore, we could say our models are robust. In overall, our methods of judgment produce good result. As the trend between normal and large dataset remain same for each algorithm, our judgment method could provide robust and consistent performance. However, there are some instances that classified incorrectly although having many same noun and verb words title and abstract, it is because the length of both title and abstract in that document are very short compared to other documents. It also failed to classify some instances that has very short title compared to its abstract. Our models are good in classifying document that has much or less same length of title and abstract to other documents in the dataset.

As the performance of LDA algorithm affected by the number of algorithm sampling iterations, our model of LDA did not performed quite good compared to the other model due to usage of small number of sampling iterations. Details about the evaluation metrics best score for normal and large dataset could be seen in Table 2 and Table 3.

Table 2. Evaluation Metrics Best Score Details, Normal Dataset

	Precision	Recall	F-Measure	Accuracy
GaN				
TF-IDF	*0.92639843*	*0.90421455*	*0.91517207*	*0.9161877*
BM25	*0.912924*	*0.95402298*	*0.93302107*	*0.931513*
LSI	*0.906542*	*0.92911877*	*0.91769158*	*0.916667*
LDA	*0.734824*	*0.88122605*	*0.8013937*	*0.781609*
ComNet				
TF-IDF	*0.97599164*	*0.93969849*	*0.95750128*	*0.9003831*
BM25	*0.96834817*	*0.9839195*	*0.97607178*	*0.975548*
LSI	*0.949153*	*0.95678392*	*0.952952953*	*0.952114*
LDA	*0.757435*	*0.81909547*	*0.78705939*	*0.793535*
Carbon				
TF-IDF	*0.96573875*	*0.94947368*	*0.95753715*	*0.9611273*
BM25	*0.9797008*	*0.96526315*	*0.97242842*	*0.974733*
LSI	*0.95186*	*0.91578947*	*0.9334764*	*0.939747*
LDA	*0.779289*	*0.78421052*	*0.78174186*	*0.797862*

Table 3. Evaluation Metrics Best Score Details

	Precision	Recall	F-Measure	Accuracy
GaN				
TF-IDF	*0.922133*	*0.9206602*	*0.9213962*	*0.921459*
BM25	*0.936047*	*0.9430244*	*0.9395225*	*0.939297*
LSI	*0.923964*	*0.9382321*	*0.9310435*	*0.930511*
LDA	*0.809147*	*0.7630457*	*0.7854207*	*0.791534*
ComNet				
TF-IDF	*0.9822335*	*0.9743202*	*0.9782608*	*0.978596*
BM25	*0.9803625*	*0.9803625*	*0.9803625*	*0.980587*
LSI	*0.927369*	*0.9707955*	*0.9485854*	*0.947984*
LDA	*0.757634*	*0.7995971*	*0.7780499*	*0.774515*
Carbon				
TF-IDF	*0.9512505*	*0.9799126*	*0.9653688*	*0.963649*
BM25	*0.9803308*	*0.9576419*	*0.9688535*	*0.968164*
LSI	*0.943572*	*0.9711790*	*0.9571766*	*0.975389*
LDA	*0.844706*	*0.7838427*	*0.8131370*	*0.813728*

IV. CONCLUSION

Our models were good in classifying document that has much or less similar length to other documents average length. As the performance metrics shown quite promising number, we inferred that our models are quite good in judging whether title matches its abstract or not. Our model also provided robust and consistent performances over different domains and size of dataset.

In the future works, we could develop into further uses. The first one is about scientific paper title evaluation. It is evaluating whether scientific paper title written nicely. It is involved syntactic and structure analysis. The second one is about scientific paper title generation, by means generate suitable title for a scientific paper that very similar to human-generated title.

ACKNOWLEDGMENT

I would also like to give thanks to all people that supported me doing the internship in Tokyo University of Agriculture and Technology. I am also very grateful to the Student Exchange Support Program (Scholarship for Short-term Study in Japan) by Japan Student Services Organization (JASSO) for making this study possible by the financial support.

REFERENCES

[1] Yucong Duan, Christophe Cruz, 2011, Formalizing Semantic of Natural Language through Conceptualization from Existence. International Journal of Innovation, Management and Technology(2011) 2 (1), pp. 37-42.

[2] Cambria, Erik; White, Bebo, 2014, "Jumping NLP Curves: A Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine 9 (2): 48–57.

[3] Christopher D. Manning, Hinrich Schütze, 1999, Foundations of Statistical Natural Language Processing, MIT Press.

[4] D. Manning, Christoper; Raghavan, Prabhakar; and Schütze, Hinrich. 2008. Introduction to Information Retrieval (online edition). Cambridge:Cambridge University Press. Page 1-18, 219-235, 403-419.

[5] Hovy, Eduard and Lin, Chin-Yew. 1999. Automated Text Summarization in SUMMARIST. TIPSTER '98 in Proceedings of a workshop on held at Baltiore, Maryland: October 13-15, 1998, pp. 197-214.

[6] Jurafsky, Daniel and H.Martin, James. 2009. Speech and Language Processing 2nd Edition. New Jersey:Person Education, Inc. Page 1-116, 759-846.

[7] Jamali, H.R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations, Scientometrics, 88 (2):653-661.

[8] Rajaraman, A.; Ullman, J. D., 2011. "Data Mining". Mining of Massive Datasets. pp. 1–17

[9] Wu, H. C.; Luk, R. W. P.; Wong, K. F.; Kwok, K. L. 2008, Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems 26 (3): 1.

[10] Le, Quoc and Mikolov, Thomas. 2014. Distributed Representations of Sentences and Documents. Mountain View:Google Inc. Page 1-4.

[11] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.