Projecting the Future Direction of Publication PatternsUsing Text Mining

Adebola K. Ojo

Department of Computer Science, University of Ibadan, Nigeria

adebola_ojo@yahoo.co.uk

Abstract - In this study, text mining techniques were used to identify various research trends in academic journal publications. These techniques were applied to figure out trends in research patterns related to various specialisation areas in Computer Science academic journal articles within a period of two decades. The corpus mined were crawled online, pre-processed and transformed into structured data using filtering and stemming algorithms. The data were grouped into series of word features based on bag of words document representation. Abstracts and the keywords of the articles selected from these journal articles were used as the dataset. It was discovered that the publication trends have changed tremendously from communications and security to artificial intelligence over time.

Index Terms– Computer Science, Filtering and Stemming Algorithms, Journals, Trends

1 INTRODUCTION

Articles published in peer reviewed journals are likely to remain a very important means of distributing research findings for the foreseeable future [1].It is a mathematical technique that uses historical results to predict future outcome. During the publication of research articles in academic journals, it is necessary to identify trends. The process of identifying trends is called the trend analysis.

Trend Analysis is a mathematical scientific approach that eliminates potential error by utilizing precise calculations in order to provide the utmost accuracy. It is the most dependable and efficient method for anticipating possible future behavior and desired outcome of a specific journal publication.It is a quantitative review of what happens over a period of time.

Article abstracts provide a comprehensive yet concise overview of an article[2][3]. Abstracts are much shorter than the full text, which minimises the influence of data noise. Therefore, this study focused on prediction of abstracts of a journal article. The data used for this study were retrieved from the Institute of Electrical and Electronics Engineers (IEEE) Transactions on Computers, a monthly publication with a wide distribution to researchers, developers, technical managers, and educators in the computer field. It publishes papers on research in areas of current interest to the readers. Journal of Institute of Electrical and Electronics Engineers (IEEE) Transactions on Computers was chosen because it is one of the highly rated Computer Science journals with ISI indexed ranking and impact factors[4][5][6][7].

This study focused on the mining of trends of journal publications using the abstracts, keywords and authors' bibliometric information of these journal articles. It was based on trend analysis using text mining techniques. Text mining is a multidisciplinary field, concerning retrieval of information, analysis of text, extraction of information, categorization, clustering, visualization, mining of data, and machine learning [8]. As the core of the knowledge management systems, text mining is a cross between information retrieval (IR) and artificial intelligence (AI). It is estimated that 90% of the world's online content is based on text (Oracle Corporation). An effective means to deal with structured, numeric content has been developed via database management systems (DBMS), but text processing and analysis is significantly more difficult. The status of knowledge management systems is much like that of DBMS twenty years ago. The real challenges, and the potential payoffs for an effective, universal text solution, are equally appealing.

Text mining predictive methods help organizations enhance the value of unstructured information by deploying insight from text analysis in software applications and business processes. Once textual information is transformed into a set

of structured data using text mining, it can be combined with traditional data mining algorithms to generate new insight for sentiment analysis and predictive analytics.

The importance of text data prediction is that, whether it is marketing and competitive intelligence, customer relationship management, social media monitoring, operational risk mitigation or threat discovery, big data is a key element for understanding where you are and where you're going. Text mining predictive methods support organizations in staying competitive. It helps them improve the ability to quickly react to customer feedback, market changes and competitive landscape evolutions.. This is precisely why enterprises should embed text analytics and predictive analytics into their business processes.

In the previous works, [9] discovered that there was percentage increase in various Computer Science disciplines such as computer architecture, artificial intelligence, scientific computing, software engineering and communication and securities. There was also a tremendous growth of computer hardware (processors, embedded systems, controllers), which occurred in 2000s.[10]was based on Trend Analysis of Machine Learning - A Text Mining and Document Clustering Methodology, by using text mining technology which collecting the homogeneous glossaries in the articles, conducting to the literature cluster analysis-based on the Social Science Citation Index (SSCI) database. [11]was based on Statistical Analysis for Monotonic Trends. The purpose of this technical note was to present and demonstrate the basic analysis of long-term water quality data for trends. This publication was targeted toward persons involved in watershed nonpoint source monitoring and evaluation projects such as those in the National Nonpoint Source Monitoring Program (NNPSMP) and the Mississippi River Basin Initiative, where documentation of water quality response to the implementation of management measures is the objective.

2 MATERIALS AND METHODS

The corpus used in IEEE consisted of the abstracts of the academic journal which were downloaded as documents from the Internet, whose volumes were published from 1997 to 2016.

Thevarious subfields in Computer Science based on Association for Computing Machinery (ACM) Computing Classification System (2016) are Algorithm and Data Structures (ADS), Artificial Intelligence/Computer vision/Robotics (AI), Communication and Security/Networking/Computer security (CS), Computer Architecture/Computer architecture/Operating systems (CA), Computer Graphics/Computer graphics/Image processing (CG), Databases/Relational databases/Data mining (DM), Programming Languages and Compilers (PLC), Scientific Computing/Bioinformatics/Computational biology (SC), Software Engineering /Human-computer interaction (SE) and, Theory of Computation/Automata theory (TOC)

One of the first steps in text or data mining is finding a way to reduce the number of independent or input variables used in the model (known as Dimensionality Reduction) without sacrificing accuracy. Dimensionality Reduction is the process of reducing the amount of variables to be used as input in a prediction model. This domain is divided into two branches: feature selection and feature extraction. Feature selection attempts to discover a subset of the original variables, while feature extraction attempts to map a high-dimensional model to a lower-dimensional space. Each word count was modified by the perceived importance of the word. Rare words carry more information than common words. TFIDF weighting of token j in document d is given as:

$$\text{tf-idf}(j, d) = \text{tf}(j, d) \times \text{idf}(j) \quad (1)$$

$$idf(j) = log\left(\frac{number\ \ of\ documents}{df(j)}\right) \quad (2)$$

tf(j, d) is the term frequency of token *j* in document *d*

df(j) is the frequency of documents containing term *j*

Term Weighting in Document Retrieval [12]

Query *q*, document *d*, with term counts *tf(j, q)* and *tf(j, d)*.

• *r(q, s):* matching score — how relevant *d* is for query *q*

• TFIDF matching score: $r(q,s) = \sum_j tf(j,q) * tf(j,d) * idf(j)$.

• TFIDF with truncated TF

$$r(q,s) = \sum_j \frac{tf(j,q).(k_3+1)}{tf(j,q)+k_3} * \frac{tf(j,d)k_1}{tf(j,d)+k_1\left((1-b)+b.\frac{l}{l}\right)} * idf(j)$$

k_1, k_3, and b are empirical constants (e.g., k1 = 1.2, k3 = 7, b = 0.75)

l/l is the document length over average document length.

3 RESULTS AND DISCUSSIONS

This section presents the summary of the feature extractions using Bags of Words representation, where each word was represented as a separate variable with its numeric weight. This consisted of all the volumes from 1997 to 2016, total number of abstracts extracted in each volume (year), the list of words generated and total words/tokens in the work. The data was text-mined and the list of words and all tokens in the work (Project) were generated for each of the years. This is presented in Table 1.

Table 1: List of Words Generated from IEEE Journal

Volume	Year	Abstracts	List of Words	Words/Tokens in the Project
IEEE 46	1997	149	3822	25852
IEEE 47	1998	138	4355	24837
IEEE 48	1999	115	3604	23132
IEEE 49	2000	123	4109	23671
IEEE 50	2001	104	5516	48713
IEEE 51	2002	124	6227	55263
IEEE 52	2003	144	7319	69037
IEEE 53	2004	131	6757	62187
IEEE 54	2005	137	6845	66961
IEEE 55	2006	141	6668	67405
IEEE 56	2007	142	7201	74521
IEEE 57	2008	135	6921	70072
IEEE 58	2009	133	8006	69219
IEEE 59	2010	140	5280	29380
IEEE 60	2011	138	5210	30914
IEEE 61	2012	145	7198	78456
IEEE 62	2013	221	7122	45151
IEEE 63	2014	248	10590	132879
IEEE 64	2015	278	11247	134286
IEEE 65	2016	290	9641	100275
TOTAL		**3,176**	**133,638**	**1,232,211**

Table 2 and Figure 1 represent total number of all word features in each of the years- for each specialisation area.

Table 2: Total Number of all Word Features in Each Year

	1997	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
ADS	327	273	246	279	272	264	263	304	297	353	378	376	281	230	262	419	368	549	567	480
AI	23	19	25	23	58	74	89	77	111	128	108	128	109	33	30	133	44	251	184	138
CS	1122	545	583	534	753	1140	1518	933	1180	1052	1340	1285	1300	747	643	1357	1210	2999	2935	1977
CA	1822	1621	1740	1617	3072	2840	3578	3331	3550	3456	3549	3390	3592	1395	1734	3912	1990	6038	5763	4975
CG	81	69	72	62	72	91	94	96	82	93	98	72	100	33	47	117	86	119	147	119
Databases	304	178	230	197	300	482	263	315	432	265	385	383	374	329	246	455	378	732	937	737
PLC	92	85	98	80	121	97	127	154	106	122	142	95	77	64	83	143	109	213	210	182
SC	107	62	49	67	153	268	259	236	252	178	312	257	272	43	63	324	110	564	736	426
SE	406	342	293	269	511	503	625	626	570	671	688	685	565	388	315	665	510	1186	1267	919
TOC	255	262	170	176	262	226	246	309	292	301	347	271	294	237	214	341	342	525	513	466
TOTAL	**6536**	**5454**	**5505**	**5304**	**7575**	**7987**	**9065**	**8385**	**8877**	**8625**	**9354**	**8950**	**8973**	**5509**	**5648**	**9878**	**7160**	**15190**	**15274**	**12435**

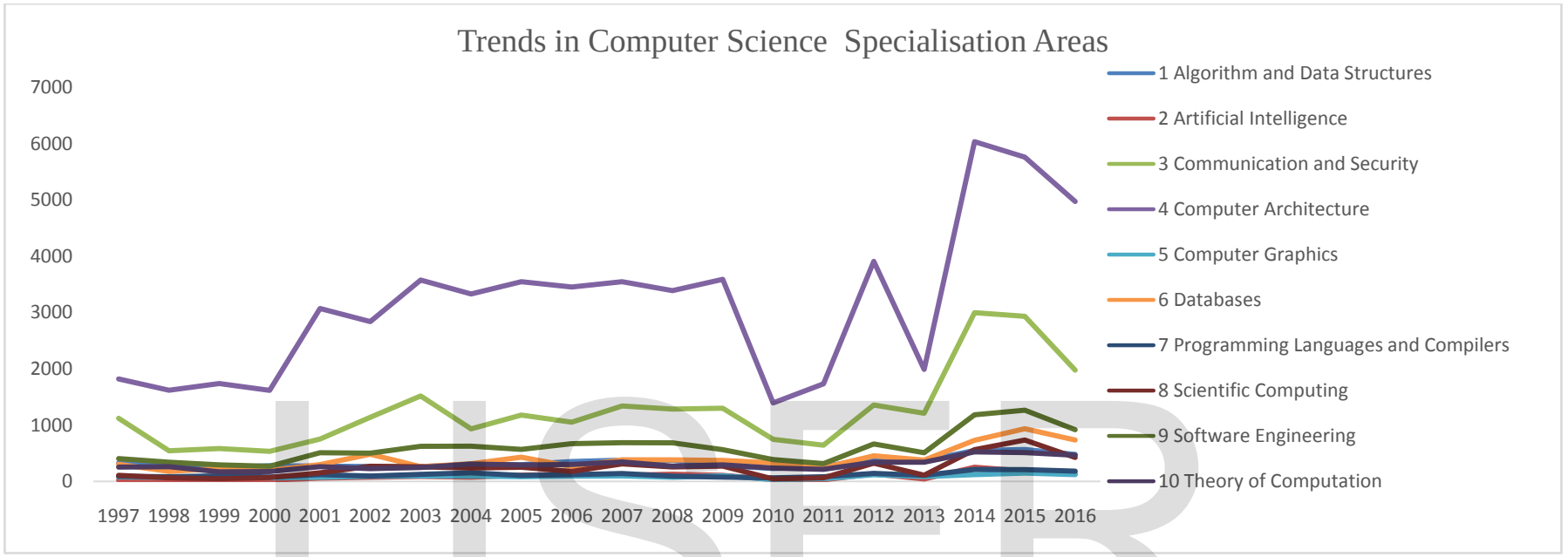

Figure 1: Publication trends from 1997 to 2016

Table 3: Classifications of IEEE Journals Based on Specialisations Areas

Specialisation Areas	1997-2001	2002-2006	2007-2011	2012-2016	Total
Algorithm and Data Structures	1397	1481	1527	2383	**6788**
Artificial Intelligence	148	479	408	750	**1785**
Communication and Security	3537	5823	5315	10478	**25153**
Computer Architecture	9872	16755	13660	22678	**62965**
Computer Graphics	356	456	350	588	**1750**
Databases	1209	1757	1717	3239	**7922**
Programming Languages and Compilers	476	606	461	857	**2400**
Scientific Computing	438	1193	947	2160	**4738**
Software Engineering	1821	2995	2641	4547	**12004**
Theory of Computation	1125	1374	1363	2187	**6049**
TOTAL	**20379**	**32919**	**28389**	**49867**	**131554**

Table 2 and Figures 2 present the classification of IEEE journals based on specialization areas in Computer Science. It comprised of all the journal volumes and their respective issues (abstracts) from 1997 (Volume 46) to 2016 (Volume 65), which was a period of 20 years. It was observed that, Computer Architecture as a specialised area has maintained a constant/steady growth in publications over the years, unlike the Artificial Intelligence, which had the highest percentage increase over time. Publications in Computer Graphics have been of low input over the years. This showed that, there has been very low turnout in journal publications in Computer Graphics, as a discipline in Computer Science.

In Table 3, for Algorithm and Data Structures, the number of its features increased gradually from 1397 to 2383 with a total of 6788 giving a 17% increase over decades Artificial Intelligence increased from 148 features to 750 totaling 1785. This gave an increase over decades of 407%. Communication and Security had 3537 initially and increased to 10478. This gave an increase of 196% over two decades. Computer Architecture increased from 9872 to 22678 making an increase of 130%. Computer Graphics increased from 356 to 588 with an increase of 65%. Databases rose to 3239 from 1209 giving a percentage increase of 168%. Programming Languages and Compilers from 476 to 857 with 80% increase in decades. Scientific Computing had its features increased from 438 to 2160 giving rise to percentage increase of 393%. In Software Engineering, the features increased from 1821 to 4547 with 150% increase over two decades. For Theory of Computation, there was an increase from 1125 to 2187, making an increase of 94% over the decades.

Table 4 presents the percentage distributions based on all areas of specialisations for each period of 5 years. The values were computed from Table 3 by diving each number of features for each specialisation, by the total number of all features of all the areas – within each period of 1997 -2001 2002 - 2006, 2007 – 2011 and 2012- 2016. This was repeated for each of the areas of specialisation.

In the first period of five-year interval (1997 and 2001), Algorithm and Data Structures was 6.9%, Artificial Intelligence: 0.7%, Communication and Security: 17.4%, Computer Architecture: 48.4%, Computer Graphics: 1.7%,Databases: 5.9%, Programming Languages and Compilers:2.3%,Scientific Computing: 2.1%, Software Engineering:8.9% and Theory of Computation:5.5%.

Between 2002 and 2006, Algorithm and Data Structures was 4.5%, Artificial Intelligence: 1.5%, Communication and Security: 17.7%, Computer Architecture: 50.9%, Computer Graphics: 1.4%, Databases: 5.3%, Programming Languages and Compilers:1.8%, Scientific Computing: 3.6%, Software Engineering:9.1% and Theory of Computation:4.2%.

Between 2007 and 2011, Algorithm and Data Structures was 5.4%, Artificial Intelligence: 1.4%, Communication and Security: 18.7%, Computer Architecture: 48.1%, Computer Graphics: 1.2%, Databases: 6.0%, Programming Languages and Compilers:1.6%, Scientific Computing: 3.3%, Software Engineering:9.3% and Theory of Computation:4.8%.

Between 2012 and 2016, Algorithm and Data Structures was 4.8%, Artificial Intelligence: 1.5%, Communication and Security: 21.0%, Computer Architecture: 45.5%, Computer Graphics: 1.2%, Databases: 6.5%, Programming Languages and Compilers:1.7%, Scientific Computing: 4.3%, Software Engineering:9.1% and Theory of Computation:4.4%.

Table 4: Percentage Distributions of Classifications of IEEE Journals

Specialisation Areas	1997-2001	2002-2006	2007-2011	2012-2016
Algorithm and Data Structures	6.9	4.5	5.4	4.8
Artificial Intelligence	0.7	1.5	1.4	1.5
Communication and Security	17.4	17.7	18.7	21.0
Computer Architecture	48.4	50.9	48.1	45.5
Computer Graphics	1.7	1.4	1.2	1.2
Databases	5.9	5.3	6.0	6.5
Programming Languages and Compilers	2.3	1.8	1.6	1.7
Scientific Computing	2.1	3.6	3.3	4.3
Software Engineering	8.9	9.1	9.3	9.1
Theory of Computation	5.5	4.2	4.8	4.4
TOTAL	**100**	100	**100**	**100**

Table 5 and Figure 2 present the percentage increase of all the areas of specialisation over two decades (within the periods of five-year intervals, that is, the trends). In Table 5, it was discovered that, Artificial Intelligence showed the highest increase of 407%. This means that, publications in this area of study increased drastically, compared to the other areas in Computer Science. Also Scientific Computing and, Communication and Security had 393% and 196% respectively. Databases, Software Engineering, Computer Architecture and Theory of Computation had maintained a steady growth of 168%, 150%, 130% and 94%. The areas which had the least growth/increase in publications were Programming Languages and Compilers, Algorithm and Data Structures and Computer Graphics. These had 80%, 71% and 65% respectively. This means that, there have been relatively low and declined percentage of publications from these areas of specialisations.

Table 5: Percentage Increase over the Two Decades

Specialisation Areas	% Increase over Decades
Algorithm and Data Structures	71
Artificial Intelligence	407
Communication and Security	196
Computer Architecture	130
Computer Graphics	65
Databases	168
Programming Languages and Compilers	80
Scientific Computing	393
Software Engineering	150
Theory of Computation	94

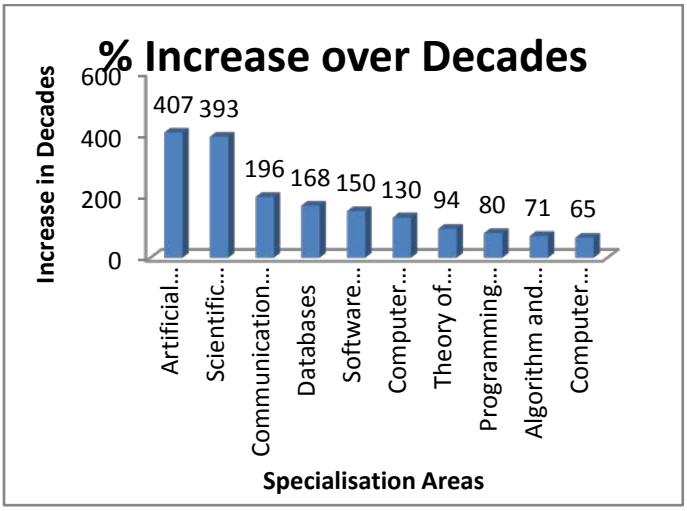

Figure 2: Percentage Increase over the Two Decades

Table 4: Percentage Distribution of Each Specialisation in Five-Year Intervals

Specialisation Areas	1997-2001	2002-2006	2007-2011	2012-2016	TOTAL
Algorithm and Data Structures	20.6	21.8	22.5	35.1	100
Artificial Intelligence	8.3	26.8	22.9	42.0	100
Communication and Security	14.1	23.2	21.1	41.7	100
Computer Architecture	15.7	26.6	21.7	36.0	100
Computer Graphics	20.3	26.1	20.0	33.6	100
Databases	15.3	22.2	21.7	40.9	100
Programming Languages and Compilers	19.8	25.3	19.2	35.7	100
Scientific Computing	9.2	25.2	20.0	45.6	100
Software Engineering	15.2	25.0	22.0	37.9	100
Theory of Computation	18.6	22.7	22.5	36.2	100

4 CONCLUSION

The corpus, which consisted of the abstracts of Journal of IEEE Transactions on Computers, one of the highly rated Computer Science journals, which were extracted/crawled as documents from the Internet. These text documents used were stored in pdf formats. These contained the list of the abstracts of the volumes published from 1997 to 2016 The subfields in Computer Science were classified based on ACM classification system into Algorithm and Data Structures, Artificial Intelligence, Communication and Security, Computer Architecture, Computer Graphics, Databases, Programming Languages and Compilers, Scientific Computing, Software Engineering, and Theory of Computation. It was observed that, Computer Architecture as a specialised area has maintained a constant growth in publications over the years.Furthermore, it was discovered that there was percentage increase over the two decades of various specialisation areas in this journal. Artificial intelligence showed the highest increase. Scientific Computing and Communication and Security, Databases, Software Engineering, Computer Architecture and Theory of Computation were the disciplines where highest percentage increase was also recorded. The areas which had the least publications were Programming Languages and Compilers, Algorithm and Data Structures and Computer Graphics respectively. Publications in Computer Graphics have been low throughout. This showed that, there has been very low turnout in journal publications from Computer Graphics, as a discipline in Computer Science.

In this study, text mining techniques were applied to figure out trends in research topics related to various specialisation areas in Computer Science academic journal articles within the period of two decades. This analysis could also be extended to find trends in research topics related to other disciplines in the academic journal articles. A similar approach can also be used to analyse academic electronic journal articles (corpus) in other fields.

5 REFERENCES

[1] Bruce A. Thyer, The Importance of Journal Articles, Oxford University Express, 2017.

[2] Carroll, Leah, "HOW TO WRITE AN ABSTRACT: Tips and Samples," pp. 1-4.

[3] Hartley, James and Cabanac, Guillaume, "ThirteenWays to Write an Abstract," *MDPI*, vol. 5, no. 11, 2017.

[4] Lowry, P. B., Moody, G. D., Gaskin, J., Galletta, D. F., Humpherys, S. L., Barlow, J. B., & Wilson, D. W., "Evaluating journal quality and the association for information systems Senior Scholars'journal basket via bibliometric measures: Do expert journal assessments add value?," *MIS Quarterly*, vol. 37, no. 4, pp. 993-1012, 2013.

[5] Ware M., Mabe, M., The STM report: an overview of scientific and scholarly journal publishing, Oxford International Association of Scientific, Technical and Medical Publishers, 2009.

[6] Andonie, R., Dzitac, I., "How to write a good paper in computer science and how will it be measured by ISI Web of Knowledge.," *International Journal of Computers, Communications & Control*, vol. 4, pp. 432-446., 2010.

[7] Ojo, Adebola K., "Predictive Analysis for Journal Abstracts using Polynomial Neural Networks

Algorithm," *IJCIS,* 2017 (Submitted for Publication).

[8] Shilpa Dang, Peerzada Hamid Ahmad , "Text Mining : Techniques and its Application," *IJETI International Journal of Engineering & Technology Innovations,* vol. 1, no. 4, November 2014.

[9] Ojo, Adebola K. and Adeyemo, Adesesan, B., "Trend Analysis in Academic Journals in Computer Science Using Text Mining," *IJCSIS,* vol. 13, no. 4, pp. 84-88, April 2015.

[10] Jiann-Min Yang ; Wei-Cheng Liao ; Wen-Chin Wu ; Chi-Yen Yin, "Trend Analysis of Machine Learning - A Text Mining And Document Clustering Methodology," *IEEC Xplore Digital Library,* 2009.

[11] Donald W. Meals, Jean Spooner, Steven A. Dressing, and Jon B. Harcum., "Statistical analysis for monotonic trends, Tech Notes 6," p. 23, 2011.

[12] Tong Zhang, "Predictive Methods for Text Mining".

[13] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Waha, David Chek Ling Ngo, "Text mining for market prediction: A systematic review," *Elsevier: Expert Systems with Applications 41 (2014)* , p. 7653–7670, 2014.

[14] Jageshwer Shriwas, Shagufta Farzana, "Using Text Mining and Rule Based Technique for Prediction of Stock Market Price," *International Journal of Emerging Technology and Advanced Engineering,* vol. 4, no. 1, pp. 246-250, 2014.

[15] Enric Junqué de Fortuny,⇈, Tom De Smedt, David Martens, Walter Daelemans, "Evaluating and understanding text-based stock price prediction," *Elsevier: Information Processing and Management* , vol. 50, p. 426–441, 2014.

[16] Yoosin Kim, Seung Ryul Jeong, Imran Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction," *Int. J. Advance. Soft Comput. Appl.,* vol. 6, no. 1, pp. 2074-8523, 2014.