

DECLARATION

I declare that the work in this thesis titled ‘Latent Dirichlet Allocation Model using Prior Derived from Empirical Data’ was carried out by me ADEGOKE, ABEJIDE MICHAEL under the supervision of Dr. A. P. Adewole, Dr. (Mrs.) F. A. Oladeji both of the Department of Computer Sciences, Faculty of Science, University of Lagos and Prof. J. O. A. Ayeni, Department of Computer Science, College of Natural Sciences, Redeemer’s University, Ede.

The information derived from literature has been duly acknowledged in the text and a list of references provided. No part of this thesis was previously presented for another degree or diploma at any other University.

.....
Student’s Name

.....
Signature

.....
Date

Dr. A. P. Adewole
(Associate Professor)
First Co-supervisor

.....
Signature

.....
Date

Dr. (Mrs.) F. A. Oladeji
(Senior Lecturer)
Second Co-supervisor

.....
Signature

.....
Date

Prof. J.O.A. Ayeni
(Professor)
Major Supervisor

.....
Signature

.....
Date

DEDICATION

This Thesis is dedicated to God, the giver of knowledge, inspiration and creativity, and to the evergreen memory of my beloved father, late Pa. Isaiah Adegoke, who did not live long enough to see the fruits of the tree he planted and scrupulously nurtured.

ACKNOWLEDGEMENTS

I am eternally grateful to my Saviour and Lord, Jesus Christ, who granted me the grace and opportunity to pursue a programme of this nature. I thank Him for His mercies, guidance, provision and faithfulness throughout the duration of the programme.

I am most grateful to my major supervisor, Prof. J. O. A. Ayeni, who not only initiated this research work, but also took time to guide, mentor and saw it to a successful completion. I equally appreciate my co-supervisors: Dr. A. P. Adewole and Dr. (Mrs.) F. A. Oladeji for their maturity, patience, tolerance and love as they, in spite of their overcrowded schedules, made out time to attend to me whenever I ran to them. I acknowledge the significant contributions of my Postgraduate teachers during the course of this programme; worthy of mention are Professor H.O.D. Longe, Dr. E.P. Fasina, Dr. B.A. Sawyerr, Dr. V.T. Odumuyiwa, Dr. O. A. Sennaike, Dr. O. B. Okunoye, Dr.(Mrs.) C.O. Yinka-Banjo and Dr. N.A. Azeez. I specially thank Professor C.O. Uwadia and Emeritus Professor O. Abass for their fatherly interest and for providing useful material during my thesis writing.

I wish to appreciate my colleagues in the Department of Computer Sciences, University of Lagos, Akoka, Nigeria: A. U. Rufai, Mrs. Afolorunsho Aderenle, Mrs. Roselyn Isimeto, Mrs. Chika Ojiakor, and Mr. Ajayi Olasupo, all whom I share views and ideas with in the course of this research. I sincerely acknowledge my employer, Bells University of Technology, Ota, Nigeria for always giving me time off my normal duty to pursue this research work. Moreover, I wish to appreciate my loving wife, Mrs. Folashade Omobola Adegoke for her support, tolerance and understanding while the programme lasted. My applause goes to my girls: Miss Favour Adegoke, Miss Testimony Adegoke and Miss Boluwatife Adegoke for their cooperation and patience during this programme. I appreciate my mother, Mrs. Lydia Ore-Ofẹ Adegoke for instilling in me rare discipline in

life and for her day-to-day prayers for me which I see as the bane of this success. My unalloyed appreciation goes to my friend and foster father, Mr. Rotimi Orowale, for his word of encouragement and prayers. Worthy of mention also is the prime support and prayers of my bosom friend, Mr. Adegboye Adegboyega. I acknowledge every other person who has been of assistance and or source of encouragement to me in one way or the other during the course of this work.

.....
Date

.....
Signature

Table of Contents

	Page
Title Page	
Declaration	i
Dedication	ii
Acknowledgements	iii
Table of Contents	v
List of Tables	viii
List of Figures	ix
Abstract	xiv
CHAPTER ONE: INTRODUCTION	
1.1 Background to the Study	1
1.2 Statement of the Problem	3
1.3 Aim and Objectives	4
1.4 Scope and Delimitation of the Study	5
1.5 Significance of the Study	5
1.6 Operational Definition of Terms	5
1.7 List of Abbreviations and Acronyms	6
CHAPTER TWO: LITERATURE REVIEW	
2.1 Text Classification Approaches	7
2.1.1 Knowledge Engineering Approach	7
2.1.2 Supervised Learning	7
2.1.3 Semi-Supervised Learning	8
2.1.4 Unsupervised Learning	10
2.2 Semantic Models for Analysis of Textual Data	11

2.2.1	Vector Space Model	12
2.2.2	Term Clustering	14
2.2.3	Latent Semantic Space Model	15
2.2.4	Probabilistic Latent Semantic Analysis	18
2.2.5	Latent Dirichlet Allocation Model	23
2.3	Related Work	31

CHAPTER THREE: METHODOLOGY

3.1	Introduction	36
3.2	Description of the formulated <i>empirical prior</i> Latent Dirichlet Allocation (<i>epLDA</i>) Model	36
3.3	Relaxation of Assumption	37
3.4	Model Formulation	39
3.5	Procedure for Implementing the Formulated Model	43
3.5.1	An Illustration of the Procedure	44
3.5.2	<i>epLDA</i> Algorithm	54
3.6	Metrics for Evaluation	57
3.6.1	Recall, Precision and F1 Measures	57
3.6.2	Perplexity Measure	58

CHAPTER FOUR: IMPLEMENTATION, RESULTS AND DISCUSSION

4.1	Data set and Experimental Setting	61
4.1.1	Data set and Data Preparation	61
4.1.2	Experimental Setting	62
4.2	System Requirements for Implementation	63
4.3	Implementation	63

4.4	Results	67
4.5	Performance Measures Using Prediction Accuracy	69
4.5.1	Model Confidence	91
4.6	Performance Measures Using Perplexity	94
4.7	Discussion	113
4.8	Summary of Findings	114
CHAPTER FIVE: CONCLUSION, CONTRIBUTIONS TO KNOWLEDGE AND FUTURE WORK		
5.1	Conclusion	115
5.2	Contributions to Knowledge	116
5.3	Future Work	116
	References	117
	List of Publications from the work	124
	Appendices	125
Appendix A	Sample documents classified into topics	125
Appendix B	Sample of Predicted and Actual Votes	133
Appendix C	Sample of Predicted	136
Appendix D	2-tailed Significant analysis between LDA and <i>ep</i> LDA	138
Appendix E	Sample Code for the <i>ep</i> LDA Program	140

List of Tables

Table	Page
1 Summary of Natural Language Text Processing Models	30
2 Summary of Related Work	34
3 Sample Cut of Set of Documents Considered	45
4 Pre-processed Documents	46
5 Term- Document Matrix Formed	48
6 U Matrix	49
7 V Matrix	50
8 Conversion of U to Probability Matrix	52
9 Conversion of V to Probability Matrix	53
10 Set of Topics Obtained Using the Prior Values U and V	54
11 Topics Obtained from the Bills of the 111th Congregational Session	67
12 Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 1 st Random Sample Using <i>ep</i> LDA Based Classifier	71
13 Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 2nd Random Sample Using <i>ep</i> LDA Based Classifier	72
14 Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 3rd Random Sample Using <i>ep</i> LDA Based Classifier	73
15 Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 4th Random Sample	

	Using <i>ep</i> LDA Based Classifier	74
16	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 5th Random Sample Using <i>ep</i> LDA Based Classifier	75
17	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 6th Random Sample Using <i>ep</i> LDA Based Classifier	76
18	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 7th Random Sample Using <i>ep</i> LDA Based Classifier	77
19	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 8th Random Sample Using <i>ep</i> LDA Based Classifier	78
20	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 9th Random Sample Using <i>ep</i> LDA Based Classifier	79
21	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 10th Random Sample Using <i>ep</i> LDA Based Classifier	80
22	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 1 st Random Sample Using LDA Based Classifier	81
23	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 2nd Random Sample Using LDA Based Classifier	82

24	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 3rd Random Sample Using LDA Based Classifier	83
25	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 4th Random Sample Using LDA Based Classifier	84
26	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 5th Random Sample Using LDA Based Classifier	85
27	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 6th Random Sample Using LDA Based Classifier	86
28	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 7th Random Sample Using LDA Based Classifier	87
29	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 8th Random Sample Using LDA Based Classifier	88
30	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 9th Random Sample Using LDA Based Classifier	89
31	Predicted Votes and % Accuracy for Congregational Sessions 111 th , 110 th and 109 th in the 10th Random Sample Using LDA Based Classifier	90

32	Average % Prediction Accuracy for each of the Congregational Sessions Using <i>epLDA</i> and LDA	91
33	Prediction Accuracy of <i>epLDA</i> on Different Training sizes	92
34	Comparison of Results with Other Models using 111 th Congregational Dataset	93
35	Collection of Documents from Associated Press data	95
36	Most Probable Words for Topics Obtained Using Traditional LDA	96
37	Most probable Words for Topics Obtained Using <i>epLDA</i>	97
38a	Perplexity Obtained for 50 Test Cases With LDA in the 1 st Random Sample	99
38b	Perplexity Obtained for 50 Test Cases With <i>epLDA</i> in the 1 st Random Sample	100
39a	Perplexity Obtained for 50 Test Cases With LDA in the 2nd Random Sample	101
39b	Perplexity Obtained for 50 Test Cases With <i>epLDA</i> in the 2nd Random Sample	102
40a	Perplexity Obtained for 50 Test Cases With LDA in the 3 rd Random Sample	104
40b	Perplexity Obtained for 50 Test Cases With <i>epLDA</i> in the 3rd Random Sample	105
137a	Perplexity Obtained for 50 Test Cases With LDA in the 100th Random Sample	106
137b	Perplexity Obtained for 50 Test Cases With <i>epLDA</i> in the 100th Random Sample	107

138	Average Perplexity for 100 Random Samples of 50 Test Cases for both
	LDA <i>ep</i> LDA in the First corpus, Second corpus and Last Corpus 108

List of Figures

Figure	Page
1 Self-Organising Map Network Architecture	9
2 Diagram of the Reduced SVD of Term-Document Matrix	17
3 PLSI as Matrix Decomposition	19
4 Statistical Inference for Topic Models	24
5 Matrix Factorisation of LDA	26
6 Graphical Plate Notation for LDA	27
7 Graphical Plate Notation for Empirical Prior Latent Dirichlet Allocation	37
8 Flow Chart for Testing <i>ep</i> LDA	54
9 Login Showing <i>ep</i> LDA Icon	65
10 Menu Screen of Visual Studio IDE	65
11 Interactive Screen Requesting For Files/data Directory	66
12 Screen Showing Directory of Files	66
13 Performance curve for Different Training sets	92
14 Comparing Results with related existing models	94
15 Comparing Average Perplexity of LDA and <i>ep</i> LDA in the first Textual Corpus	111
16 Comparing Average Perplexity of LDA and <i>ep</i> LDA in the second Textual Corpus	112
17 Comparing Average Perplexity of LDA and <i>ep</i> LDA in the last Textual Corpus	113

Abstract

One of the problems confronting the full use of the power of the computers is that they understand very little of the meaning of human language. Significant progress is therefore being made to develop computational tools that will help organise data (text corpora) that will support computer users to quickly find relevant information from the sea of collective knowledge that are digitised and stored in online databases. Many semantic models, including latent Dirichlet allocation, have been proposed to help computers to deal with the potential vagueness that may arise due to variability in word usage. Latent Dirichlet allocation model reduces each document in a text collection to a mixture of topics that summarises the main themes in the collection. The Latent Dirichlet Allocation (LDA) model has been widely used to identify and extract hidden structures in data. Recent literature, however, reported that the model suffers from the restriction that the values of its controlling parameters, namely, prior distributions for the computation of the mixture components for theme extractions are not derived from data. Rather, pre-allocated, fixed priors are adopted and used irrespective of domain of application. The use of pre-allocated priors is based on the assumption that the computation of thematic structures is independent of the occurrence of words and documents in text collections. This assumption is, however, too strong and it has been observed that usage of pre-allocated priors which are often not consistent with the underlying data has led to some well-developed models failing to produce reasonable predictions in real application. In this study, *empirical prior* latent Dirichlet allocation (*epLDA*) model that uses latent semantic indexing framework to derive the priors required for topics computation from data is presented. The derived priors incorporate knowledge from the data into the LDA model. The parameters of the priors so obtained are related to the parameters of the conventional LDA model using exponential function. The model was implemented using C# programming language and tested on benchmarked data. It achieved higher prediction accuracy than the conventional latent Dirichlet allocation (LDA), supervised latent Dirichlet allocation (sLDA) and other existing models that have used the same data set for predictive tasks. It was observed that the *epLDA* model consistently outperforms the conventional LDA on different datasets; its performance falls within highly sure confidence level. The best known reported model in literature, Random Walk Heterogeneous Graph (RWHG), achieves a prediction accuracy of 90.36 percent while the proposed model achieves a prediction accuracy of 92.15 percent thereby providing higher prediction confidence. The model also achieves lower perplexity resulting in better generalisation performance than the conventional LDA model on the same dataset. The average generalisation performance of the model on test data is 65.46 while that of the conventional LDA on the same dataset is 72.94.

Keywords: latent Dirichlet allocation; semantic indexing; empirical priors; hidden structures; perplexity measure.