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	……..……	…….. Date
Dr. (Mrs.) F. A. Oladeji (Senior Lecturer) Second Co-supervisor	……..……	…….. Date
Prof. J.O.A. Ayeni (Professor) Major Supervisor	……..……	…….. 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-Ofe 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

List of Tables

List of Figures

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 (*ep*LDA) 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 *ep*LDA 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.