AN ENHANCED SEQUENTIAL EXCEPTION TECHNIQUE FOR SEMANTIC-BASED TEXT ANOMALY DETECTION

MOHAMMED AHMED TAIYE

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2019

AN ENHANCED SEQUENTIAL EXCEPTION TECHNIQUE FOR SEMANTIC-BASED TEXT ANOMALY DETECTION

By

MOHAMMED AHMED TAIYE

Thesis Submitted to

Awang Had Salleh Graduate School of Arts and Sciences,

Universiti Utara Malaysia

In Fulfilment of the Requirement for the Degree of Doctor of Philosophy

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

MOHAMMED AHMED TAIYE

calon untuk Ijazah
(candidate for the degree of) **PhD**

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"AN ENHANCED SEQUENTIAL EXCEPTION TECHNIQUE FOR SEMANTIC-BASED TEXT ANOMALY DETECTION"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***02 Mei 2019.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: ***May 02, 2019.***

Pengerusi Viva: *(Chairman for VIVA)*	**Prof. Dr. Huda Hj Ibrahim**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Assoc. Prof. Dr. Shuzlina Abdul Rahman**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Dr. Mohamad Farhan Mohamad Mohsin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof. Dr. Siti Sakira Kamaruddin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Dr. Farzana Kabir Ahmad**	Tandatangan *(Signature)*

Tarikh:
(Date) ***May 02, 2019***

Permission to Use

In presenting this study in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this study in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in her absent, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this study or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my study.

Requests for permission to copy or to make other use of materials in this study, in whole or in part should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Acknowledgments

All praise to Almighty Allah (SWT) who gave me courage and patience to carry out this work. Alhamdulillah.

My sincere appreciation and gratitude goes to my supervisors, Professor Madya Dr. Siti Sakira Kamurddin and Dr. Farzana Kabir Ahmed. My appreciation also goes to my scholarship guarantor and lecturer Dr. Norliza Binti Katuk and My Late supervisor Professor Mohammed Syazwan B. Abdullah for their academic guidance, support and encouragement. Not forgotten the appointed examiners who have given valuable comments to improve my study.

I wish to express my gratitude to the academic and supporting staff of School of Computing, Universiti Utara Malaysia for all the assistance rendered during my studies. I wish to thank all my friends, whose continuous discussions and, support greatly helped in this research.

I would like to dedicate this work to my twin brother Mohammed Kehinde Mohammed and my son Mohammed Nabeel, thank you. My wife Maryam Olaoti Shehu Mohammed for the patience and support. my caring elder siblings Ibrahim Mohammed and my Lovely sister Maryam Aiyelero Mohammed thank you for being there for me, especially when I needed you most Jazaka Allah kheir. My Parents Mr. Aliyu Jimoh Mohammed and Mrs. Racheal Aliyu Mohammed for always being their Jazaka Allah kheir for your love and support during my difficult times. my sincere appreciation goes to my parent-in-law, Mr and Mrs Shehu Olaoti Jazaka Allah kheir for your support and encouragement during my period of study I am grateful. Lastly, my relatives, colleagues, friends and football teammates. You all made my academic journey worthwhile. Thank you

Abstrak

Pengesanan anomali teks berasaskan semantik adalah bidang penyelidikan yang menarik dan telah mendapat perhatian daripada komuniti perlombongan data. Pengesanan anomali teks mengenal pasti maklumat yang menyimpang daripada maklumat am yang terkandung dalam dokumen. Data teks dikaitkan dengan masalah kekaburan, keamatan tinggi, bersela dan perwakilan teks. Sekiranya cabaran ini tidak diselesaikan dengan baik, pengenalpastian anomali teks berasaskan semantik akan menjadi kurang tepat. Kajian ini mencadangkan Teknik Pengecualian Jujukan yang ditambah baik (ESET) untuk mengesan anomali teks berasaskan semantik dengan mencapai lima objektif: (1) untuk mengubahsuai Teknik Pengecualian Jujukan (SET) dalam memproses teks tidak berstruktur; (2) untuk mengoptimumkan Kesamaan Kosain bagi mengenal pasti data teks serupa dan tidak serupa; (3) untuk menghibridkan SET yang diubahsuai dengan Analisis Semantik Laten (LSA); (4) untuk mengintegrasikan algoritma Lesk dan Pemilihan Keutamaan bagi penyahtaksaan makna dan mengenal pasti bentuk kanonik teks; dan (5) untuk mewakili anomali teks berasaskan semantik menggunakan Logik Tertib Pertama (FOL) dan Graf Konsep Rangkaian (CNG). ESET melaksanakan pengesanan anomali teks dengan menggunakan Kesamaan Kosain yang dioptimumkan, menghibridkan LSA dengan SET yang diubahsuai, dan mengintegrasikannya dengan algoritma Penyahtaksaan Makna Perkataan khususnya Lesk dan Pemilihan Keutamaan. Kemudian, FOL dan CNG dicadangkan untuk mewakili anomali teks berasaskan semantik yang dikesan. Bagi menunjukkan ketersauran teknik tersebut, empat set data telah dipilih untuk diuji iaitu data NIPS, ENRON, blog Daily Koss, dan 20Newsgroups. Penilaian eksperimen menunjukkan ESET telah meningkatkan ketepatan pengesanan anomali teks berasaskan semantik daripada dokumen. Apabila dibandingkan dengan pengukuran sedia ada, keputusan eksperimen telah mengatasi kaedah penanda aras dengan skor F1 yang lebih baik daripada semua set data; Data NIPS 0.75, ENRON 0.82, blog Daily Koss 0.93 dan 20Newsgroups 0.97. Hasil yang dijana daripada ESET telah terbukti signifikan dan menyokong tanggapan yang semakin berkembang mengenai anomali teks berasaskan semantik dalam literatur yang sedia ada. Secara praktikal, kajian ini menyumbang kepada pemodelan topik dan pertautan konsep bagi tujuan menggambarkan maklumat, perkongsian pengetahuan dan mengoptimumkan pembuatan keputusan.

Kata Kunci: Kesamaan semantik, Anomali teks berasaskan semantik, Penyahtaksaan Makna Perkataan, Teknik Pengecualian Jujukan ditambah baik.

Abstract

The detection of semantic-based text anomaly is an interesting research area which has gained considerable attention from the data mining community. Text anomaly detection identifies deviating information from general information contained in documents. Text data are characterized by having problems related to ambiguity, high dimensionality, sparsity and text representation. If these challenges are not properly resolved, identifying semantic-based text anomaly will be less accurate. This study proposes an Enhanced Sequential Exception Technique (ESET) to detect semantic-based text anomaly by achieving five objectives: (1) to modify Sequential Exception Technique (SET) in processing unstructured text; (2) to optimize Cosine Similarity for identifying similar and dissimilar text data; (3) to hybridize modified SET with Latent Semantic Analysis (LSA); (4) to integrate Lesk and Selectional Preference algorithms for disambiguating senses and identifying text canonical form; and (5) to represent semantic-based text anomaly using First Order Logic (FOL) and Concept Network Graph (CNG). ESET performs text anomaly detection by employing optimized Cosine Similarity, hybridizing LSA with modified SET, and integrating it with Word Sense Disambiguation algorithms specifically Lesk and Selectional Preference. Then, FOL and CNG are proposed to represent the detected semantic-based text anomaly. To demonstrate the feasibility of the technique, four selected datasets namely NIPS data, ENRON, Daily Koss blog, and 20Newsgroups were experimented on. The experimental evaluation revealed that ESET has significantly improved the accuracy of detecting semantic-based text anomaly from documents. When compared with existing measures, the experimental results outperformed benchmarked methods with an improved F1-score from all datasets respectively; NIPS data 0.75, ENRON 0.82, Daily Koss blog 0.93 and 20Newsgroups 0.97. The results generated from ESET has proven to be significant and supported a growing notion of semantic-based text anomaly which is increasingly evident in existing literatures. Practically, this study contributes to topic modelling and concept coherence for the purpose of visualizing information, knowledge sharing and optimized decision making.

Keywords: Semantic similarity, Semantic-based text anomaly, Word Sense Disambiguation, Enhanced Sequential Exception Technique.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
BCD	Block Coordinate Descent
CG	Conceptual Graphs
CGIF	Conceptual Graph Interchange Format
CNG	Concept Network Graph
ESET	Enhanced Sequential Exception Technique
FCA	Formal Concept Analysis
FCA-RS	Similarity measure proposed in Wang and Liu
FOL	First Order Logic
GSDPMM	Gibbs Sampling algorithm for Dirichlet Multinomial Mixture Model
GMM	Gaussian Mixture Model
HDP	Hierarchical Dirichlet Processing
HMM	Hidden Markov Model
k-NN	k-Nearest Neighbour
LCH	Leacock & Chodorow
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
LSA	Latent Semantic Analysis
MMR	Maximum Marginal Relevance
NED	News Event Detection
NER	Name Entity Recognition
NG	Network Graph
NIPS	Neural Information Processing Systems
NLP	Natural Language Processing
NMF	Non-Negative Matrix Factorization
NMI	Normalized Mutual Information
OLAP	Online Analytical Processing

PCA	Principal Component Analysis
PMI-IR	Point wise Mutual Information using data collected
PLSA	Probabilistic Latent Semantic Analysis
PLSI	Probabilistic Latent Semantic Index
POI	Persons of Interest
POS	Part of Speech
RES	Resnik
SET	Sequential Exception Technique
SVD	Singular Value Decomposition
SVM	Support Vector Machine
WUP	Wu & Palmer
WSD	Word Sense Disambiguation

Definition of Terms

Anomaly	observation which deviates so much from other observations as to arouse uncertainties that it was produced by an alternate mechanism
Algorithm	set of rules to be followed in solving a problem.
Semantic	meaning relating to words and phrases.
Corpus	collection of written texts, especially the entire works of an author or a body of writing on a subject.
Count vectorization	transformation of text into vector representations so that numeric machine learning approach such as counting can be done easily.
Disambiguation	removal of vagueness or ambiguity by making a context understandable or clear in meaning.
Dissimilarity	difference or variance
Cardinality	total count of elements present in a set or group, as a property of that group.
Measure	process of ascertaining the size or degree of an object.
Method	procedure or an approach of accomplishing a task.
Modification	the process of changing or adapting to improve an object
Network Graph	vertices or nodes that are connected by edges.
LAS	co-occurred terms found in corpus are captured using dimensionality reduction approach (SVD) on a term-by-document matrix *T* representing corpus
FOL	computational approach to knowledge representation following the language rules of grammatical representation.

Pruning	process of reducing the complexities of classifiers and hence improving its accuracy by reducing overfits.
SET	sequential exception technique recreates an approach in which unusual objects can be differentiated from series of like objects.
SVD	Singular Value decomposition is used to simplify or ease term vectorization in Text mining
Technique	way of carrying out an operation

CHAPTER ONE

INTRODUCTION

1.1 Overview

This study presents an Enhanced Sequential Exception Technique (ESET) for semantic-based text anomaly detection. The study focuses on enhancing a technique that gives a better detection accuracy in identifying and representing semantic-based text anomalies in documents. To achieve this, chapter one was structured as thus; Section 1.2 briefly discuss the study research background. Section 1.3 states the research problem. Section 1.4 outlines the research question. Section 1.5 outlines the research objectives. Section 1.6 presents the research scope. Section 1.7 presents significance of the study and section 1.8 presents the organization of thesis.

1.2 Research Background

Enhanced sequential exception technique was used in this study to detect semantic based text anomaly in documents. Hence, various unique methods have emerged over the years to satisfy the need of detecting semantic based text anomaly (Arning & Rakesh, 1996; Kamaruddin, 2011;. Kamaruddin et al., 2015; Kamaruddin, Hamdan, Bakar, & Mat Nor, 2012; Takahashi, 2011; Upadhyaya & Singh, 2012). With advancement in technology, the overload phenomenon of text document needs to be properly managed for knowledge sharing purposes and optimized decision making (Lee, et.al, 2017). Text information is one of the most valuable assets in the world today. Nonetheless, discovering meaningful knowledge from large volume of text document is tasking (Debortoli, Müller, Junglas, &

The contents of the thesis is for internal user only

References

A.Rajaraman, J. Leskovec, J. D. U. (2016). *Mining Massive Data Sets Winter 2016*. Cambridge University Press. Retrieved from http://web.stanford.edu/class/cs246

ABDULSAHIB, A. K. (2015). *Graph based text representation for document clustering asma khazaal abdulsahib*.

Abdulsahib, A. K., & Kamaruddin, S. S. (2015). Graph based text representation for document clustering. *Journal of Theoretical and Applied Information Technology*, *76*(1), 1–13. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84930694414&partnerID=40&md5=5c7f0059c26594915cdf9360315173c7

Abouzakhar, N., Allison, B., & Guthrie, L. (2008). Unsupervised Learning-based Anomalous Arabic Text Detection. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, 291–296. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/summaries/83.html

Acree, B., Jansa, J., & Shoub, K. (2016). Comparing and Evaluating Cosine Similarity Scores, Weighted Cosine Similarity Scores, and Substring Matching. Retrieved from https://shoub.web.unc.edu/files/2016/04/AHJS_Weighted_Cosine.pdf

Adler-Golden, S. M. (2009). Improved hyperspectral anomaly detection in heavy-tailed backgrounds. *WHISPERS '09 - 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2–5. https://doi.org/10.1109/WHISPERS.2009.5289019

Aggarwal, C., & Zhai, C. (2012). *Mining text data*. (C. C. C. Z. AGGARWAL, Ed.), *Mining Text Data* (Vol. 4). Kluwer Academic Publishers Boston/Dordrecht/London. https://doi.org/10.1007/978-1-4614-3223-4

Agirre, E., & Martinez, D. (2002). Integrating selectional preferences in WordNet. *Proceedings of the First International WordNet Conference*, 9. Retrieved from http://arxiv.org/abs/cs/0204027

Akarsu, B., Bayram, K., Slisko, J., & Corona Cruz, A. (2013). International Journal Of Scientific Research And Education. *Ijsae.In*, *6*(3), 221–232. Retrieved from http://ijsae.in/ijsaeems/index.php/ijsae/article/viewFile/157/137

Akoglu, L., Tong, H., & Koutra, D. (2014). Graph-based Anomaly Detection and Description: A Survey. *ArXiv Preprint ArXiv:1404.4679*, 49. https://doi.org/10.1007/s10618-014-0365-y

Alagi, D. (2009). Experiments on Active Learning for Croatian Word Sense Disambiguation.

Allan Collins, J. S. B., Larkin, & K. M., & Newman, B. B. and. (2007). *INFERENCE IN TEXT UNDERSTANDING*. University of Illinois at Urbana- Champaign 51 Gerty Drive Champaign, Illinois 61820.

Allan, J., Carbonell, J., & Doddington, G. (1998). Topic detection and tracking pilot study: Final report. *DARPA Broadcast News Transcription and Understanding Workshop.*, 194–218. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.6373&rep=rep1

&type=pdf

Almarimi, A., & Andrejková, G. (2016). Text Anomalies Detection Using Histograms of Words. *ACSIJ Advances in Computer Science: An International Journal*, *5*(1), 63–68.

Arning, A., & Rakesh, A. (1996). Method for Deviation in Large Databases. *KDD-96 Proceedings*.

Atefeh, F., & Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter TECHNIQUES FOR EVENT DETECTION IN TWITTER. *Computational Intelligence*, *0*(1), 132–164. https://doi.org/10.1111/coin.12017

Balbi, S. (2010). Beyond the curse of multidimentionality: high dimentional clustering in context mining. *Statistica Applicata - Italian Journal of Applied Statistics*, *22*(1), 53–63.

Banerjee, S. (2002). Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet, (December).

Basile, P., Caputo, A., & Semeraro, G. (2014). An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 14)*, 1591–1600.

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications*, *91*, 159–169. https://doi.org/10.1016/j.eswa.2017.08.047

Beltagy, I., Roller, S., Cheng, P., Erk, K., & Mooney, R. J. (2015). Representing Meaning with a Combination of Logical Form and Vectors, 1–44. Retrieved from http://arxiv.org/abs/1505.06816

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic Parsing on Freebase from Question-Answer Pairs. *Proceedings of EMNLP*, (October), 1533–1544. Retrieved from https://www.aclweb.org/anthology/D/D13/D13-1160.pdf%5Cnhttp://www.samstyle.tk/index.pl/00/http/nlp.stanford.edu/pubs/semparseEMNLP13.pdf

Bernotas, M., Karklius, K., Laurutis, R., & Slotkiene, A. (2007). The peculiarities of the text document representation, using ontology and tagging-based clustering technique. *Information Technology and Control*, *36*(2), 217–220.

Bertoldi, N., Cettolo, M., & Federico, M. (2010). Statistical Machine Translation of Texts with Misspelled Words. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, (June), 412–419.

Bhaduri, K., Matthews, B. L., & Giannella, C. R. (2011). Algorithms for speeding up distance-based outlier detection. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 859–867. https://doi.org/10.1145/2020408.2020554

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2012). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Boyd-Graber, J., Blei, D. M., & Zhu, X. (2007). A Topic Model for Word Sense Disambiguation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, 1024–1033.

Brants, T., Chen, F., & Farahat, A. (2003). A system for new event detection. *ACM SIGIR Conference on Research and Development in Informaion Retrieva*, (pp. 330-337).

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-based document segmentation with probabilistic latent semantic analysis. *Proceedings of the Eleventh International Conference on Information and Knowledge Management CIKM 02*, 211. https://doi.org/10.1145/584792.584829

Breja, M. (2015). A Novel approach for Novelty Detection of Web Documents, *6*(5), 4257–4262.

Brody, S. (2005). Cluster-Based Pattern Recognition in Natural Language Text. *English*, (August). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.81.7288&rep=rep1&type=pdf

Bruynooghe, M., & Denecker, M. (2014). First Order Logic with Inductive Definitions for Model-Based Problem Solving.

Bustince, H., Fernadez, J., & Mesiar, R. (2011). Restricted dissimilarity functions and penalty functions. *Eusflat-Lfa 2011*, (July). Retrieved from

http://library.utia.cas.cz/separaty/2012/E/mesiar-restricted dissimilarity functions and penalty functions.pdf

Cai, D., He, X., Wu, X., & Han, J. (2008). Non-negative matrix factorization on manifold. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 63–72. https://doi.org/10.1109/ICDM.2008.57

Cambria, E., & Melfi, G. (2015). Semantic Outlier Detection for Affective Common-Sense Reasoning and Concept-Level Sentiment Analysis, 276–281.

Cammert, M., Heinz, C., Kramer, J., & Riemenschneider, T. (n.d.). Systems and/or methods for event stream deviation detection. *U.S. Patent No. 9,659,063. Washington, DC: U.S. Patent and Trademark Office*. Retrieved from https://www.google.com/patents/US9659063

Capurro, I., Lecumberry, F., Martín, Á., Ramírez, I., Rovira, E., & Seroussi, G. (2016). Efficient sequential compression of multi-channel biomedical signals. *IEEE Journal of Biomedical and Health Informatics*, *PP*(NN), 13. Retrieved from http://arxiv.org/abs/1605.04418

Cha, S. (2007). Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions, *1*(4).

Chandarana, D. R. (2015). A Survey for Different Approaches of Outlier Detection in Data Mining, 1–4.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(September), 1–58.

https://doi.org/10.1145/1541880.1541882

Chaplot, D. S., & Salakhutdinov, R. (2018). Knowledge-based Word Sense Disambiguation using Topic Models. Retrieved from http://arxiv.org/abs/1801.01900

Chen, X., & Wu, C. (2012). A Text Representation Method Based on Harmonic Series. In *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012* (pp. 1830–1834).

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, I. (2008). Text classification and Naive Bayes. Retrieved from lp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html

Cichosz, P. (2018). Anomaly detection in discussion forum posts using global vectors. In *SPIE*. Proc. SPIE 10808. https://doi.org/10.1117/12.2501345

Classen, A., Boucher, Q., & Heymans, P. (2011). A text-based approach to feature modelling: Syntax and semantics of TVL. *Science of Computer Programming*, *76*(12), 1130–1143. https://doi.org/10.1016/j.scico.2010.10.005

Dang, S., & Ahmad, P. H. (2014). Text Mining : Techniques and its Application, *1*(4), 22–25.

Debortoli, S., Müller, O., Junglas, I. A., & vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Tutorial. *Manuscript Submitted for Publication*, (April).

Deshpande, R., Vaze, K., Rathod, S., & Jarhad, T. (2014). Comparative Study of Document Similarity Algorithms and Clustering Algorithms for Sentiment Analysis. *Ijettcs.Org*, *3*(5), 196–199. Retrieved from http://www.ijettcs.org/Volume3Issue5/IJETTCS-2014-10-21-85.pdf

Ding, R., Nallapati, R., Xiang, B., & Services, A. W. (2016). Coherence-Aware Neural Topic Modeling, *1*.

Drissi, M., & Watkins, O. (2017). Hierarchical Text Generation using an Outline.

Eshghi, A., Howes, C., Gregoromichelaki, E., Hough, J., & Purver, M. (2015). *Feedback in Conversation as Incremental Semantic Update*. *Iwcs 2015*. Retrieved from http://www.aclweb.org/website/old_anthology/W/W15/W15-01.pdf#page=123

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. https://doi.org/10.18653/v1/W16-2506

Foltz, P. W. (1996). Latent Semantic Analysis for Text-Based. *Behavior Research Methods, Instruments and Computers*, *28*(2), 197–202. https://doi.org/10.3758/BF03204765

Franzoni, V. (2017). Just an Update on PMING Distance for Web-based Semantic Similarity in Artificial Intelligence and Data Mining, 1–3. https://doi.org/10.13140/RG.2.2.20531.22560

Froud, H., Lachkar, A., & Ouatik, S. (2013). Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. *ArXiv Preprint*

ArXiv:1302.1612. Retrieved from http://arxiv.org/abs/1302.1612

Furtado, P., Nadal, S., Peralta, V., Djedaini, M., & Marcel, P. (2015). Materializing Baseline Views for Deviation Detection Exploratory OLAP, 1–12.

Fyshe, A., Talukdar, P., Murphy, B., & Mitchell, T. (2013). Documents and Dependencies : an Exploration of Vector Space Models for Semantic Composition. *Conll*, 84–93.

Gabrilovich, Evgeniy, and S. M. (2005). Feature generation for text categorization using world knowledge. *IJCAI International Joint Conference on Artificial Intelligence*, *5*(pp. 1048-1053.).

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJCAI International Joint Conference on Artificial Intelligence*, 1606–1611. https://doi.org/10.1145/2063576.2063865

Gahl, S., Menn, L., Ramsberger, G., Jurafsky, D. S., Elder, E., Rewega, M., & Audrey, L. H. (2003). Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences. *Brain and Cognition*, *53*(2), 223–228. https://doi.org/10.1016/S0278-2626(03)00114-3

Garrette, D., Erk, K., & Mooney, R. (2014). A Formal Approach to Linking Logical Form and Vector-Space Lexical Semantics. *Computing Meaning SE - 3*, *47*, 27–48. https://doi.org/10.1007/978-94-007-7284-7_3

Gelbukh, A., Sidorov, G., & Han, S.-Y. (2005). On some optimization heuristics for lesk-like WSD algorithms. *Nldb'05*, 402–405.

Giannoulis, P., Potamianos, G., & Maragos, P. (2018). On the Joint Use of NMF and Classification for Overlapping Acoustic Event Detection. *Proceedings*, *2*(2), 90. https://doi.org/10.3390/proceedings2020090

Gilad Katz, Yuval Elovici, & B. S. (2014). *SEMANTIC BASED CONTEXTUAL CLUSTERING FOR DATA LEAKAGE PREVENTION THROUGH ANOMALY DETECTION*.

Gloor, P. A., Niepel, S., L, Y., Whalley, G., Skilling, J. K., Kitchen, L., & Causey, R. (2006). Identifying Potential Suspects by Temporal Link Analysis Discovering Suspicious Activity in the Enron e-Mail Dataset Filtering by Keywords, 9.

Godbole, S. (2002). Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers. *Progress Report, IIT Bombay*, (March 2002), 17. Retrieved from http://www.it.iitb.ac.in/~shantanu/work/report.pdf

Goldstein, M., Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLoS ONE*, (April), 1–31. https://doi.org/10.7910/DVN/OPQMVF

Gomaa, W. H. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, *68*(13), 13–18.

Gong, Y., Zhao, K., & Zhu, K. Q. (2016). Representing Verbs as Argument Concepts. *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*, 2615–2621.

Goodfellow, I. (2016). NIPS 2016 Tutorial: Generative Adversarial Networks. https://doi.org/10.1001/jamainternmed.2016.8245

Guthrie, D. (2008). Unsupervised Detection of Anomalous Text. *Distribution*, (July).

Guthrie, D., Guthrie, L., Allison, B., & Wilks, Y. (2007). Unsupervised anomaly detection. *IJCAI International Joint Conference on Artificial Intelligence*, 1624–1628.

H, S. D., M, M. K., & Science, C. (2015). International Journal of Combined Research & Development (IJCRD) eISSN : 2321-225X ; pISSN : 2321-2241 Volume : 4 ; Issue : 2 ; February -2015 A Survey on Text Mining Approaches International Journal of Combined Research & Development (IJCRD), 251–256.

Han, J. (2014). Data Mining : Concepts and Techniques.

Hardin, J. S., Sarkis, G., & Urc, P. C. (2015). Network analysis with the enron email corpus. *Journal of Statistics Education*, *23*(2). https://doi.org/10.1080/10691898.2015.11889734

Hassan, S., & Mihalcea, R. (2011). Semantic Relatedness Using Salient Semantic Analysis. *Proceedings of the 25th AAAI Conference on Artificial Intelligence, (AAAI 2011)*, 884–889. Retrieved from http://www.samerhassan.com/images/4/48/Hassan.pdf%5Cnhttp://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/download/3616/3972

Héas, P., Drémeau, A., & Herzet, C. (2016). An Efficient Algorithm for Video Superresolution Based on a Sequential Model. *SIAM Journal on Imaging Sciences*,

9(2), 537–572. https://doi.org/10.1137/15M1023956

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravičius, V., & Duneld, M. (2014). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, *5*(1), 6. https://doi.org/10.1186/2041-1480-5-6

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodoligies. *Artificial Intelligence Review*, *22*(1969), 85–126. https://doi.org/10.1007/s10462-004-4304-y

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, (April), 49–56. Retrieved from http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

Issa, H., & Vasarhelyi, M. A. (2011). Application of Anomaly Detection Techniques to Identify Fraudulent Refunds. *SSRN Working Papers Series*, 1–19. https://doi.org/10.2139/ssrn.1910468

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Janz, A., Kędzia, P., & Piasecki, M. (2018). Graph-based complex representation in inter-sentence relation recognition in Polish texts. *Cybernetics and Information*

Technologies, *18*(1), 152–170. https://doi.org/10.2478/cait-2018-0013

Jiang, L., Zhang, H., Yang, X., & Xie, N. (2013). Research on Semantic Text Mining Based on Domain Ontology, 336–343.

Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning*, 137–142. https://doi.org/10.1007/BFb0026683

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*, *21*, 0–934. https://doi.org/10.1162/089120100750105975

Kamaruddin, S. S. B. (2011). *FRAMEWORK FOR DEVIATION DETECTION IN TEXT*.

Kamaruddin, S. S., Bakar, A. A., Hamdan, A. R., Nor, F. M., Nazri, M. Z. A., Othman, Z. A., & Hussein, G. S. (2015). A text mining system for deviation detection in financial documents. *Intelligent Data Analysis*, *19*(s1), S19–S44. https://doi.org/10.3233/IDA-150768

Kamaruddin, S. S., Hamdan, A. R., & Bakar, A. A. (2007). Text Mining for Deviation Detection in Financial Statement, 446–449.

Kamaruddin, S. S., Hamdan, A. R., Bakar, A. A., & Mat Nor, F. (2012). Deviation detection in text using conceptual graph interchange format and error tolerance dissimilarity function. *Intelligent Data Analysis*, *16*(3), 487–511.

https://doi.org/10.3233/IDA-2012-0535

Kamruzzaman, S. M., Haider, F., & Hasan, A. R. (2010). Text Classification using Data Mining. *Science*, 19. Retrieved from http://arxiv.org/abs/1009.4987

Kannan, R., Woo, H., Aggarwal, C. C., & Park, H. (2017). Outlier Detection for Text Data : An Extended Version. *ArXiv*, 489–497.

Kannan, Ramakrishnan, Woo, H., Aggarwal, C. C., & Park, H. (2017). Outlier Detection for Text Data : An Extended Version. Retrieved from http://arxiv.org/abs/1701.01325

Karkali, M., Rousseau, F., Ntoulas, A., & Vazirgiannis, M. (2014). Using temporal IDF for efficient novelty detection in text streams. *ArXiv*, 30. Retrieved from http://arxiv.org/abs/1401.1456

Katariya, N. P., & Chaudhari, M. S. (2015). 126. Text Preprocessing for Text Mining Using Side Information. *International Journal of Computer Science and Mobile Applications*, *3*, 3–7.

Kim, J., & Montague, P. (2017). An Efficient Semi-Supervised SVM for Anomaly Detection, 2843–2850.

Kobus, C., Yvon, F., & Damnati, G. (2008). Normalizing SMS: are two metaphors better than one? *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, (August), 441–448. Retrieved from http://dl.acm.org/citation.cfm?id=1599137

Koehrsen, W. (2017). Machine Learning with Python on the Enron Dataset. Retrieved November 23, 2018, from https://medium.com/@williamkoehrsen/machine-learning-with-python-on-the-enron-dataset-8d71015be26d

Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J., Smith, N. a, & Dyer, C. (2015). Frame-Semantic Role Labeling with Heterogeneous Annotations. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 218–224.

Kumar, a A. (2012). Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering Sri Sivani College of Engineering Sri Sivani College of Engineering, *1*(5), 1–6.

Kumar Palaniswamy Supervisor, H., & Aldous, D. (2015). Exploratory Data Analysis of Enron Emails.

Kumaraswamy, R., & Shavlik, J. (2012). Anomaly Detection in Text : The Value of Domain Knowledge, 225–228.

Lee, Hanjun, Keunho Choi, Donghee Yoo, Yongmoo Suh, Soowon Lee, G. H. (2017). Recommending valuable ideas in an open innovation community A text mining approach to information overload problem. https://doi.org/10.1108/eb057530

Lenci, A., Montemagni, S., & Pirrelli, V. (2001). The Acquisition and Representation of Word Meaning The Acquisition and Representation of Word Meaning . An Overview.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries. *Proceedings of the 5th Annual International Conference on Systems Documentation - SIGDOC '86*, 24–26. https://doi.org/10.1145/318723.318728

Leveling, J. (2007). IRSAW – Towards Semantic Annotation of Documents for Question Answering.

Leyzerov, O. (2017). Identifing Fraud from Enron Email and financial data. Retrieved November 23, 2018, from https://olegleyz.github.io/enron_classifier.html

Li, L., Hu, X., Hu, B. Y., Wang, J., & Zhou, Y. M. (2009). Measuring sentence similarity from different aspects. *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*, *4*(July), 2244–2249. https://doi.org/10.1109/ICMLC.2009.5212182

Li, L. I. N., Hu, X. I. A., Hu, B., Wang, J. U. N., & Zhou, Y. (2009). MEASURING SENTENCE SIMILARITY FROM DIFFERENT ASPECTS, (July), 12–15.

Li, X., Member, D. F., Croft, W. B., Head, D., & University, B. E. T. (2006). Sentence Level Information Patterns for Novelty Detection, 1–10. https://doi.org/10.1145/1183614.1183652

Liang, H., Tsai, F. S., & Kwee, A. T. (2009). Detecting novel business blogs. *ICICS 2009 - Conference Proceedings of the 7th International Conference on Information, Communications and Signal Processing*. https://doi.org/10.1109/ICICS.2009.5397541

Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2014). A Similarity Measure for Text

Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, *26*(7), 1575–1590. https://doi.org/10.1109/TKDE.2013.19

Liu, H., Ke, W., Wei, K. K., & Hua, Z. (2013). The impact of IT capabilities on firm performance: The mediating roles of absorptive capacity and supply chain agility. *Decision Support Systems*, *54*(3), 1452–1462. https://doi.org/10.1016/j.dss.2012.12.016

Liu, Z. (2013). *High Performance Latent Dirichlet Allocation for Text Mining*.

M. J. Denny & A. Spirling. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.

Mahapatra, A., Srivastava, N., & Srivastava, J. (2012). Contextual anomaly detection in text data. *Algorithms*, *5*(4), 469–489. https://doi.org/10.3390/a5040469

Maitra, Anutosh (Bangalore, I., Mohamedrasheed, Annervaz Karukapadath (Trichur, I., Jain, Tom Geo (Bangalore, I., Shivaram, Madhura (Bangalore, I., Sengupta, Shubhashis (Bangalore, I., Ramnani, Roshni Ramesh (Bangalore, I., … Sahu, Vedamati (Bangalore, I. (2016). SYSTEM FOR AUTOMATED ANALYSIS OF CLINICAL TEXT FOR PHARMACOVIGILANCE. Retrieved June 17, 2016, from http://www.freepatentsonline.com/y2016/0048655.html

Manevitz, L. M. (2001). One-Class SVMs for Document Classification. *Journal of Machine Learning Research*, *2*, 139–154. https://doi.org/10.1162/15324430260185574

Margaret Rouse. (2005). First order predicate Logic. Retrieved October 3, 2015, from

http://whatis.techtarget.com/definition/first-order-logic

Marvin, R. (2018). Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems, *1*, 125–131.

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, *46*(6), 1116–1124. https://doi.org/10.1016/j.jbi.2013.08.008

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *IMIA Yearbook of Medical Informatics Methods Inf Med*, *47*(1), 128–144. https://doi.org/me08010128

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. https://doi.org/10.1.1.65.3690

Miller, R. C., & Myers, B. A. (2001). Outlier finding. *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology - UIST '01*, 81. https://doi.org/10.1145/502348.502361

Montes-y-gómez, M., Gelbukh, A. F., & López-lópez, A. (2002a). Detecting Deviations in Text Collections: An Approach Using Conceptual Graphs. *Mexican International Conference on Artificial Intelligence*, 176–184. https://doi.org/10.1007/3-540-46016-0_19

Montes-y-gómez, M., Gelbukh, A., & López-lópez, A. (2002b). Text Mining at Detail

Level Using Conceptual Graphs, 122–136.

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, *19*(03), 291–330. https://doi.org/10.1017/S1351324913000065

Navigli, R. (2009a). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 10. https://doi.org/10.1145/1459352.1459355

Navigli, R. (2009b). Word sense disambiguation. *ACM Computing Surveys*, *41*(2), 1–69. https://doi.org/10.1145/1459352.1459355

Ngai, E. W. T., Hong, T., Polytechnic, K., Hom, H., Kong, H., Hom, H., & Kong, H. (2016). a Review of the Literature on Applications of Text Mining in Policy Making.

Oberreuter, G., & Velásquez, J. D. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications*, *40*(9), 3756–3763. https://doi.org/10.1016/j.eswa.2012.12.082

Otterbacher, J., & Radev, D. (2006). Fact-focused novelty detection: A feasibility study. *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *2006*, 687–688. https://doi.org/10.1145/1148170.1148318

Pappas, Y. (2018). Fraud Detection Using Machine Learning (Analysis). Retrieved November 23, 2018, from http://www.yannispappas.com/Fraud-Detection-Using-Machine-Learning/

Parr, T. (2012). jguru. Retrieved January 1, 2015, from http://www.jguru.com/faq/view.jsp?EID=81

Patel, F. N., & Soni, N. R. (2012). Text mining: A Brief survey. *International Journal of Advanced Computer Research*, *2*(6), 243–248. Retrieved from http://www.theaccents.org/ijacr/papers/conference/icett2012/43.pdf

Pawar, A. M. (2015). A Comprehensive Survey on Online Anomaly Detection, *119*(17), 41–45.

Peter Norvig. (2015). Natural Language Processing What We Do. Retrieved December 9, 2015, from http://research.google.com/pubs/NaturalLanguageProcessing.html

Poon, H., & Domingos, P. (2010). Unsupervised ontology induction from text. *Proceedings of the 48th Annual Meeting of the …*, (July), 296–305. Retrieved from http://dl.acm.org/citation.cfm?id=1858712

Powers, D. M. W. (2015). What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. https://doi.org/KIT-14-001

Pradhan, N., Gyanchandani, M., & Wadhvani, R. (2015). A Review on Text Similarity Technique used in IR and its Application. *International Journal of Computer Applications*, *120*(9), 29–34. https://doi.org/10.5120/21257-4109

Provost, F., Fawcett, T., & Kohavi, R. (1997). The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning1*, 445–453.

Ramage, D., Heymann, P., Manning, C. D., & Garcia-Molina, H. (2009). Clustering the tagged web. *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, 54. https://doi.org/10.1145/1498759.1498809

Ramya, R. S., Venugopal, K. R., Iyengar, S. S., & Patnaik, L. M. (2016). Feature Extraction and Duplicate Detection for, *16*(5).

Ray, S., & Craven, M. (2001). Representing sentence structure in hidden Markov models for information extraction. *International Joint Conference On*, *17*(1), 1273–1279. Retrieved from http://scholar.google.com/scholar?q=intitle:Representing+Sentence+Structure+in+Hidden+Markov+Models+for+Information+Extraction#0

Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. *Information Sciences*, *236*, 109–125. https://doi.org/10.1016/j.ins.2013.02.029

Rennie, J. (2008). 20 Newsgroups. Retrieved November 2, 2018, from http://qwone.com/~jason/20Newsgroups/

Rosario, B., & Hearst, M. a. (2004). Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 430. https://doi.org/10.3115/1218955.1219010

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the Joint Conference on*

Empirical Methods in Natural Language Processing and Computational Natural Language (EMNLP-CoNLL'07), *1*(June), 410–420. https://doi.org/10.7916/D80V8N84

Rumshisky, A. (2008). Resolving Polysemy in Verbs: Contextualized Distributional Approach to Argument Semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, Special Issue of Italian Journal of Linguistics*, 1–27.

Sardar, R. P. ; S. S. ; S. K. N. ; M. M. (2018). Improving Lesk by Incorporating Priority for Word Sense Disambiguation. https://doi.org/10.1109/EAIT.2018.8470436

Sayeed, A., Greenberg, C., & Demberg, V. (2016). Thematic fit evaluation: an aspect of selectional preferences. *ACL 2016, 99.*

Silveira, S. B., & Branco, A. (2012). Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. *Proceedings of the 2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012*, (1), 482–489. https://doi.org/10.1109/IRI.2012.6303047

Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications*, *80*(10), 25–33. https://doi.org/10.5120/13897-1851

Steinberger, J., & Ježek, K. (2004). Using Latent Semantic Analysis in Text Summarization. *In Proceedings of ISIM 2004*, 93--100.

Sugiyama, M., & Borgwardt, K. (2013). Rapid Distance-Based Outlier Detection via

Sampling. *Advances in Nueral Information Processing Systems 26 (Proceedings of NIPS)*, 1–9.

Sun, F., Guo, J., Lan, Y., Xu, J., & Cheng, X. (2016). Semantic Regularities in Document Representations. Retrieved from http://arxiv.org/abs/1603.07603

Szmeja, P., Ganzha, M., Paprzycki, M., & Pawłowski, W. (2018). Dimensions of Semantic Similarity, 87–125.

Takahashi, T. (2011). Discovering Emerging Topics in Social Streams via Link Anomaly Detection.pdf, *26*, 1–18. https://doi.org/10.1109/icdm.2011.53

Tan, L., Zhang, H., Clarke, C. L. a, & Smucker, M. D. (2015). Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. *Acl*, 657–661.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. *Introduction to Data Mining*, 769.

Torres, S., & Gelbukh, A. (2009). Comparing Similarity Measures for Original WSD Lesk Algorithm. *Advances in Computer Science and Applications*, *43*, 155–166.

Tsai, F. S. (2007). Novelty detection for text documents using named entity recognition. *2007 6th International Conference on Information, Communications & Signal Processing*, (3), 1–5. https://doi.org/10.1109/ICICS.2007.4449883

Turney, P. D., & Pantel, P. (2010). ★★★★★From Frequency to Meaning_ Vector Space Models of Semantics（讲的非常好，但是我还只看了三分之一）.pdf, *37*,

141–188. https://doi.org/10.1613/jair.2934

Upadhyaya, S., & Singh, K. (2012). Classification based outlier detection techniques. *Int J Comput Trends Technol*, *3*, 294–298. Retrieved from http://www.ijcttjournal.org/Volume3/issue-2/IJCTT-V3I2P118.pdf

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. *Proceedings of ECAI Workshop on Ontology Learning and Population*, 37–42. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Enriching+a+Lexical+Semantic+Net+with+Selectional+Preferences+by+Means+of+Statistical+Corpus+Analysis#0

Wang, Y., Ni, X., Sun, J.-T., Tong, Y., & Chen, Z. (2011). Representing document as dependency graph for document clustering. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11*, 2177. https://doi.org/10.1145/2063576.2063920

Wehmeier, K. F. (2004). Wittgensteinian Predicate Logic. *Notre Dame Journal of Formal Logic*, *45*(1), 1–11. https://doi.org/10.1305/ndjfl/1094155275

William Wei Song, Chenlu Lin, A. F. (2017). An Euclidean similarity measurement approach for hotel rating data analysis. Retrieved from https://ieeexplore.ieee.org/abstract/document/7951927/authors

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *WWW '13 Proceedings of the 22nd International Conference on World Wide Web*,

1445–1456. Retrieved from http://dl.acm.org/citation.cfm?id=2488388.2488514

Yang, Y., Zhang, J., Carbonell, J., & Jin, C. (2002). Topic-conditioned novelty detection. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '02*, 688. https://doi.org/10.1145/775047.775150

Yih, W., & Meek, C. (n.d.). Improving Similarity Measures for Short Segments of Text, 1489–1494.

Yin, J., & Wang, J. (2016). A Model-based Approach for Text Clustering with Outlier Detection. *Icde*, 625–636. https://doi.org/10.1109/ICDE.2016.7498276

Yoo, J., & Yang, D. (2015). Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier Text Classification using TF-IDF and Naï ve Bayes Classifier, *111*(Comcoms), 263–266. https://doi.org/10.14257/astl.2015.111.50

Yuhanis, S. S. kamaruddin and Y. (2015). constructing canonical data model for text document clustering, 4.

Zhang, D., Zhai, C., Han, J., Srivastava, A., & Oza, N. (2009). Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications. *Statistical Analysis and Data Mining*, *2*(5–6), 378–395. https://doi.org/10.1002/sam.10059

Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to TExt classification using Semi-supervised Clustering. *Knowledge-Based Systems*, *75*, 152–160.

https://doi.org/10.1016/j.knosys.2014.11.028

Zhang, W., Xiao, F., Li, B., & Zhang, S. (2016). Using SVD on Clusters to Improve Precision of Interdocument Similarity Measure. *Computational Intelligence and Neuroscience*, *2016*. https://doi.org/10.1155/2016/1096271

Zhang, Z. Z. Z., & Feng, X. F. X. (2009). New Methods for Deviation-Based Outlier Detection in Large Database. *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, *1*. https://doi.org/10.1109/FSKD.2009.303

Zhou, G., Zhao, J., Liu, K., & Cai, L. (2011). Exploiting Web-Derived Selectional Preference to Improve Statistical Dependency Parsing. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1556–1565.

Zhou, Y., Fleischmann, K. R., & Wallace, W. A. (2010). Automatic text analysis of values in the enron email dataset: Clustering a social network using the value patterns of actors. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1–10. https://doi.org/10.1109/HICSS.2010.77

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577. https://doi.org/ROC; Receiver-Operating Characteristic; SDT; Signal Detection Theory

LIST OF PUBLICATIONS WITH RESEARCH & GRANT WORK

2015 A Framework For Semantic-based Anomaly Detection In Text, 4th International Conference on Internet Applications, Protocols and Services*(NETAPP)*, Malaysia December 1-3, 2015.
2015 Expert Directory System for Managing Organizational Knowledge
2016 Representing Semantics of Text By Acquiring Its Canonical Form, 3rd International Multi-conference on Artificial Intelligence Technology *(M-CAIT 2016)*, Bangi, Selangor Malaysia August 23-24, 2016.
2017 Representing Semantics of Text by Acquiring its Canonical Form
2017 Framework for Enhancing A Wearable Device that Converts Sound, Text and Image Into Automatic Sign Language Recognizing System (ASLR)
2017 Framework on comparative analysis of Text Representation Schemes and Similarity Measures For Sentences
2017 Combined Word Sense Disambiguation Algorithms with Latent Semantic Analysis to identify semantic similarity in unstructured textual data
2017 Visualization of Spoken Language for Deaf People
2018 Graph-based Representation for Sentence Similarity Measure: A Comparative Analysis

List of papers in-view

- Extraction of Agro-food terms from online news website in Malaysia
- Integration of Word sense disambiguation algorithms to analyze and identify similar terms in documents
- Optimizing sequential exception techniques for anomaly detection in corpuse
- A systematic review on text anomaly (from all levels of text; word, sentence, document, events and topics)

APPENDIX A

Process Flow in ESET

As described in section 3.3 This study centres on the enhancement of ESET following this flow.

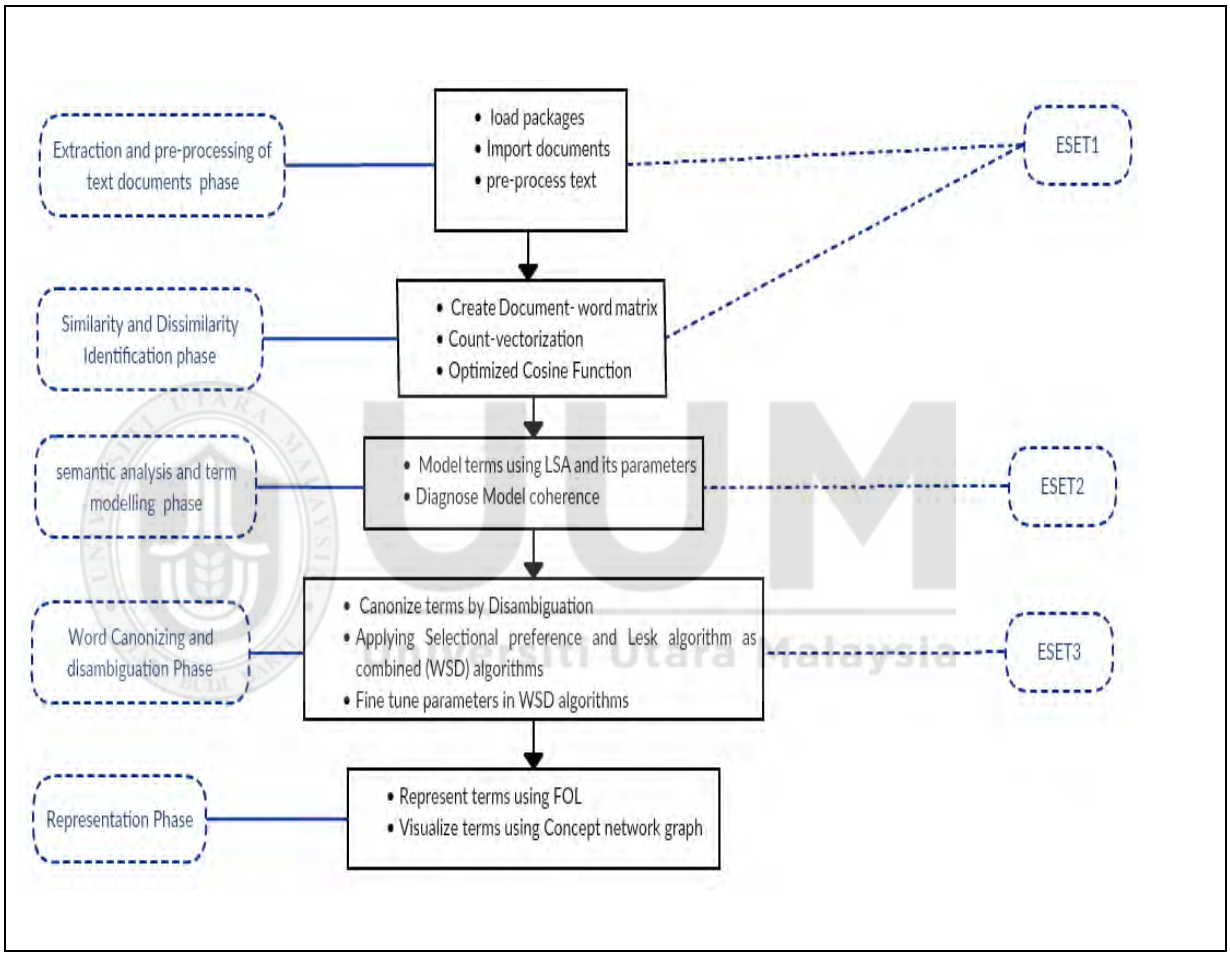

Figure A1. ESET Flowchart

APPENDIX B

Code Snippet of Results Extracted from ENRON POI

As described in section 4.2 in chapter four an experiment was carried out to assess the feasibility of the ESET. In this experiment, sentences were extracted and tested from Mail messages sent and received from Kenneth Lay. Figure A1 shows why and how Kenneth Lay was chosen as a POI.

Data Preparation Phase

```
# ### Importing libraries and magics
# ### Import the file which contain the data to our variable
# In[3]:
# Load the dictionary containing the data
with open(os.getcwd()+"/final_project_data.pkl", "rb") as data_file:
    data_init = pickle.load(data_file)
# ### Converting the data from a python dictionary to a pandas dataframe
# In[4]:
#Converting the data from a python dictionary to a pandas dataframe
data_df = pd.DataFrame.from_dict(data_init, orient='index')
raw_data = data_df.copy()
# #### Now check the structure of the new data frame to find out how
many total number of observation and column are prsent
# In[5]:
data_df.count().sort_values()
# In[9]:
#dropping 'poi' and 'email_address' variables
data_df = data_df.drop(["email_address"], axis=1)
data_temp = data_df.drop(["poi"], axis=1)
data_temp[data_temp.isnull().all(axis=1)]
        ys = dataframe[[feature]]
        quartile_1, quartile_3 = np.percentile(ys, [25, 75])
        iqr = quartile_3 - quartile_1
        lower_bound = int(round(quartile_1 - (iqr * 3)))
        upper_bound = int(round(quartile_3 + (iqr * 3)))
        partial_result = list(np.where((ys > upper_bound) | (ys <
lower_bound))[0])
        print(feature, len(partial_result))
        result.update(partial_result)
    print("Total number of records with extreme values: " +
    selector = SelectPercentile(percentile=100)
    a = selector.fit(X, y)
    plt.figure(figsize=(12,9))
    sns.barplot(y=X.columns, x=a.scores_)
# In[40]:
plot_importance(data)
```

SET Phase

```
# coding: utf-8
# ### Sequential Exception Technique (SET)
# Identify the POIs using SET and print their names.
SET_data.head()
# In[46]:
cols = [ 'salary', 'bonus', 'long_term_incentive', 'deferred_income',
def SET(m,SET_data):
# Set the value of parameter m = the no. of iterations you require
    Card = pd.Series(np.NAN)
    DS=pd.Series(np.NAN)
    idx_added = pd.Series(np.NAN)
    pos = 0
    for j in range(1,m+1):
        new_indices                                                    =
np.random.choice(e_names.index,len(e_names),replace=False)
        for i in pd.Series(new_indices).index:
            idx_added[i+pos] = new_indices[i]
 DS[i+pos]=sum(np.var(SET_data.loc[e_names[new_indices[:i+1]]]))
            Card[i+pos] = len(e_names[:i+1])
        pos = pos+i+1
    df = pd.DataFrame({'Index_added':idx_added,'DS':DS,'Card':Card})
    df ['DS_Prev'] = df.DS.shift(1)
    df['Card_prev'] = df.Card.shift(1)
    df.Card_prev[(df.Card == 1)] = 0
    df = df.fillna(0)
    df['Smoothing'] = (df.Card - df.Card_prev)*(df.DS - df.DS_Prev)
    # find indexes of sets with max sf
    maxsf = []
    for i in range(len(df.DS)):
        if df.Smoothing[i] == df.Smoothing.max():
            maxsf.append(i)
    #print(maxsf)
    N = len(e_names)
    excp_set = []
    for i in range(len(maxsf)):
        j = maxsf[i]
        k=j+1
        temp = []
        temp.append(df.Index_added[j])
        excp_set.append(temp.copy())
        temp_prev = pd.DataFrame()
        temp_j = pd.DataFrame()
        a=j
        while(a%N!=0):
            temp_row = SET_data.loc[e_names[df.Index_added[a]]]
            temp_j = temp_j.append(temp_row)
            a=a-1
        temp_row = SET_data.loc[e_names[df.Index_added[a]]]
        temp_j = temp_j.append(temp_row)
        temp_prev = temp_j.copy()                # Ij-1
```

```
        temp_prev.drop(temp_prev.index[0],inplace=True)
        #temp_prev.index = np.arange(len(temp_prev))
        while(k%N!=0):
            K_element = SET_data.loc[e_names[df.Index_added[k]]]   # K th
element
            temp_prev = temp_prev.append(K_element)          # Ij-1 U {ik}
            temp_j = temp_j.append(K_element)             # Ij U {ik}
            Dk0 = sum(np.var(temp_prev)) - df.DS[j-1]
            Dk1 = sum(np.var(temp_j)) - df.DS[j]
            if Dk0-Dk1 >= df.DS[j]:                    # If Dk0 - Dk1 >= Dj
excp_set[i].append(df.Index_added[k])
            temp_prev.drop(temp_prev.index[len(temp_prev)-
1],inplace=True)
            temp_j.drop(temp_j.index[len(temp_j)-1],inplace=True)
            k+=1
    #print(excp_set)                          # contains the indices of exception
elements.
    return excp_set
# In[ ]:
excp_set = SET(1000,SET_data)
# In[ ]:
# Printing the POIs.
print("\nException set: \n")
for i in range(len(excp_set)):
    print(e_names[excp_set[i]])
```

Figure A1 code snippet of POI with the most Total Payment Information using SET.

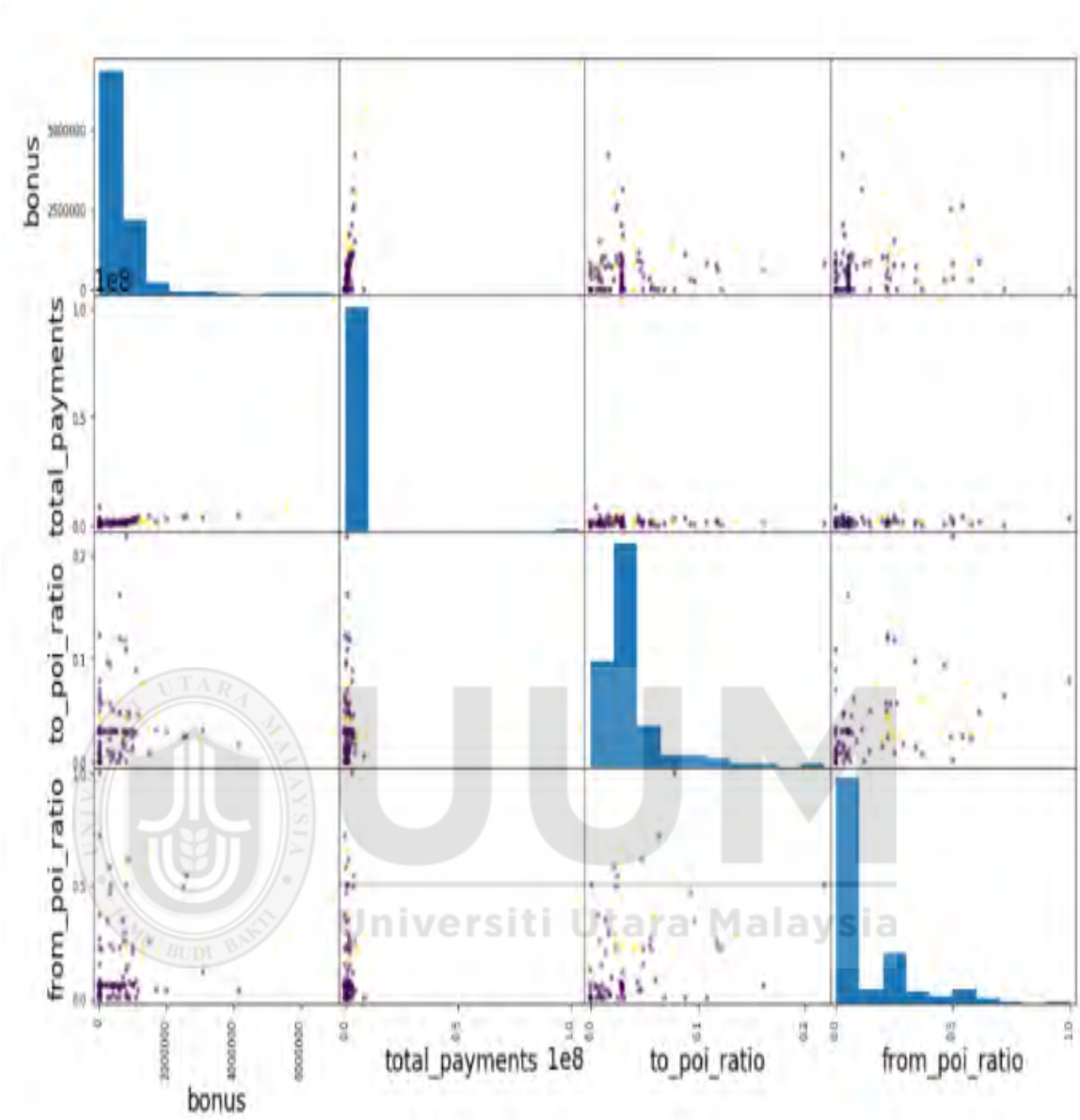

Figure B1 Data pre-processing Phase

Figure A2 shows a scatter matrix with an overall visualization of the ENRON email financial data.

OUTPUT	
FREVERT MARK A	12
BELDEN TIMOTHY N	9
SKILLING JEFFREY K	9
BAXTER JOHN C	8
LAVORATO JOHN J	8
DELAINEY DAVID W	7
KEAN STEVEN J	7
HAEDICKE MARK E	7
WHALLEY LAWRENCE G	7
RICE KENNETH D	6
KITCHEN LOUISE	6
LAY KENNETH L	**15**

Figure A3 presents the Output of SET codes.

After identifying the most POI (Kenneth Lay) by comparing Total payment information as seen in the figures shown above. To identify other POI who mail and received messages from Kenneth Lay, some pseudocodes were also developed to identify other POIs. In addition, pseudocodes used in extracted mail messages from the identified POI.

Analysing ENRON and Identifying most sent and received Mails

```
import os
from collections import Counter
from email.parser import Parser
rootdir = "C:\\Users\\Shantnu\\Desktop\\Data Sources\\maildir\\"
def email_analyse(inputfile, to_email_list, from_email_list,
email_body):
    with open(inputfile, "r") as f:
        data = f.read()
    email = Parser().parsestr(data)
    if email['to']:
        email_to = email['to']
        email_to = email_to.replace("\n", "")
        email_to = email_to.replace("\t", "")
        email_to = email_to.replace(" ", "")
        email_to = email_to.split(",")
        for email_to_1 in email_to:
            to_email_list.append(email_to_1)
    from_email_list.append(email['from'])
to_email_list = []
from_email_list = []
email_body = []
for directory, subdirectory, filenames in  os.walk(rootdir):
    for filename in filenames:
```

Word Frequency

```
        email_analyse(os.path.join(directory, filename), to_email_list,
from_email_list, email_body )
print("\nTo email adresses: \n")
print(Counter(to_email_list).most_common(10))
print("\nFrom email adresses: \n")
print(Counter(from_email_list).most_common(10))
import os
from collections import Counter
from email.parser import Parser
rootdir = "C:\\Users\\Shantnu\\Desktop\\Data Sources\\maildir\\"
def email_analyse(inputfile, to_email_list, from_email_list,
email_body):
    with open(inputfile, "r") as f:
        data = f.read()
    email = Parser().parsestr(data)
    if email['to']:
        email_to = email['to']
        email_to = email_to.replace("\n", "")
        email_to = email_to.replace("\t", "")
        email_to = email_to.replace(" ", "")
        email_to = email_to.split(",")
        for email_to_1 in email_to:
            to_email_list.append(email_to_1)
    from_email_list.append(email['from'])
to_email_list = []
from_email_list = []
email_body = []
for directory, subdirectory, filenames in  os.walk(rootdir):
    for filename in filenames:
        email_analyse(os.path.join(directory, filename), to_email_list,
from_email_list, email_body )
print("\nTo email adresses: \n")
print(Counter(to_email_list).most_common(10))
print("\nFrom email adresses: \n")
print(Counter(from_email_list).most_common(10))
```

Figure A4 presents a code snippet for extracting and analysing mail messages sent and received from POIs

Mail messages of POIs were all analysed and extracted. These mail messages contain other attributes like senders and recipients ID dates and the mime version. In this study, we are only interested in extracting the body of mail messages to avoid unnecessary content and as well reduce the workload of text data pre-processing.

APPENDIX C

A Sample of Most Frequent Terms Using ESET

No	Term1	Term2	Term 3	Term 4
1	Enron	Time	Please	Deal
2	Business	Thank	Thank	Gas
3	manage	Day	Attach	Price
4	Meet	Don't	Email	Contract
5	Market	Call	Enron	Power
6	Company	Talk	Call	Rate
7	Vince	Hope	Copying	Trade
8	Report	Ill	Fax	Day
9	Time	Bit	File	Month
10	Energy	Trying	Message	Companies
11	Information	Guy	Information	Energy
12	Please	Night	Phone	Transaction
13	Trade	Friday	Send	Product
14	Discuss	Weekend	Corp	Term
15	Regards	Love	Kay	Custom
16	Team	Item	Receive	Cost
17	Plan	Email	Question	Thank
18	Service	people	Draft	Purchase
19	Message	File	Price	Organization
20	Phone	Information	Business	Electricity

Figure A5 presents list of ENRON Terms

Figure A5 is a list of terms from ENRON using the ESET2 of the study research design. Daily Kos Bloggs. The study also presents list of some terms from the Daily Kos blogs data using the word-cloud as a visualization scheme