KNOWLEDGE BASED SEMANTIC REPRESENTATION FOR SEMANTIC RELATEDNESS MEASUREMENTS

ALI MUTTALEB HASAN

DOCTOR OF PHILOSOPHY

UNIVERSITI MALAYSIA PAHANG

SUPERVISOR'S DECLARATION

We hereby declare that We have checked this thesis, and, in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy.

(Supervisor's Signature)

Full Name : TS. DR. TAHA HUSSEIN ALAALDEEN RASSEM

Position : SENIOR LECTURER

Date : 26/10/2022

(Co-supervisor's Signature)

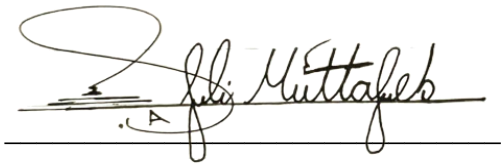
Full Name : TS. DR. NOORHUZAIMI@KARIMA BINTI MOHD NOOR

Position : SENIOR LECTURER

Date : 26/10/2022

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.



(Student's Signature)

Full Name : ALI MUTTALEB HASAN

ID Number : PCC17001

Date : 26/10/2022

KNOWLEDGE BASED SEMANTIC REPRESENTATION FOR SEMANTIC RELATEDNESS MEASUREMENTS

ALI MUTTALEB HASAN

Thesis submitted in fulfillment of the requirements
for the award of the Best thesis of the degree of
Doctor of Philosophy

Faculty of Computing
UNIVERSITI MALAYSIA PAHANG

OCTOBER 2022

ACKNOWLEDGEMENTS

First of all, I would like to thank the supreme power the Almighty God who is obviously the one has always guided me to work on the right path of my life. Without his grace this project could not became a reality.

It would not have been possible to write this doctoral thesis without the help and support of the kind people around me, to only some of whom it is possible to give a particular mention here. Above all, I would like to thank my beloved parents, my father and mother, for their personal support and very great patience at all times. This thesis is a gift to them, including my brother and sisters, who have given me their unequivocal support throughout, as always, for which my mere expression of thanks, likewise, does not suffice.

This thesis would not have been thinkable without help, support, and patience of my principal supervisors, (Ts. Dr. Taha Hussein Alaaldeen Rassem and Co-supervisor, Ts. Dr. Noorhuzaimi@Karimah Binti Mohd Noor). Their good advice, encouragement, and friendship have been invaluable on both an academic and personal level, for which I am grateful.

I would like to acknowledge the financial, academic, and technical support of the University Malaysia Pahang as they gave me a very good lap for the study of my research journey and a special thanks to our faculty of computing staff for doing their best for us. This research, particularly in the award of a postgraduate research fellowship as funded by RDU, PGRS and SPPS, provided the necessary financial support for this research. The library facilities and computer facilities of the university, as well as the National Library of University Malaysia Pahang, for their support and assistance since the start of my postgraduate work from 2016 to 2021.

I am highly obliged to take this opportunity to sincerely thank all the staff members of the computer department for their generous attitude and friendly behavior.

Last but not least, I am grateful to all of the UMP and IPS staff who have assisted and encouraged me throughout the year. I have no valuable words to express my thanks, but my heart is still full of the favors received from every person.

ABSTRAK

Textual analysis has become one of the most important tasks due to the rapid increase in the number of texts. The text has been continuously generated in a variety of formats, including social media postings and chats, emails, articles, and news. The handling of these texts necessitates efficient and effective procedures capable of dealing with linguistic challenges arising from natural language complexity. In recent years, there has been a lot of research into using semantic characteristics from lexical sources to deal with synonymy and ambiguity difficulties in text mining tasks like document clustering and classification. The main challenges of exploiting the lexical knowledge sources included WordNet in how to incorporate the different types of semantic relations for capturing more semantic evidence and how to handle the high dimensionality of the current semantic representation approaches. The research proposes a new knowledge-based semantic representation approach for semantic relatedness measurements. The weighting-based method for incorporating the semantic relations in the lexical sources is proposed to form the representation vector of the word. The proposed approach depends on the topological parameters (depth, density, descendants, and ancestors) in the semantic taxonomy. To handle the high dimensionality issue in the weighting-based method, a new topic-based technique is introduced to represent the semantics of words in terms of topics instead of the concepts in the weighting-based method. This proposed approach depends on the semantic features in the lexical sources (such as WordNet) for handling the synonymy and ambiguity issues. The proposed approach is evaluated for semantic relatedness measurements using six gold standard test sets. The evaluation results in terms of correlation measures demonstrate that the weighting-based method is more effective than the state-of-the-art feature-based methods. For the sample's harmonic measure to be accurate, the most anomalous value of r and p is calculated using the measure of the mean for each dataset, the proposed *r* and *p* methods are MC30, RG65, WordRel353, MT287, MEN3000, and Rgnew65 *r* 0.82, 0.86, 0.52, 0.53, 0.89, and 0.47, also for *p* 0.80, 0.82, 0.52, 0.47, 0.82, and 0.45. The results of the measurements indicated from the datasets are measures of the standard Means, thus the results of measurements of the proposed approach are 0.81, 0.84, 0.46, 0.49, 0.52, and 0.86 for MC30, RG65, WordRel353, MT287, MEN3000, and Rgnew65, respectively. The Non-zero is utilised to assess the proposed approach in order to ascertain the percentage of word pairings with a semantic relatedness value larger than zero. Using MC30, RG65, WordRel35, MT287, MEN3000, and Rgnew65, the NZ attained in the experiments was 0.96, 0.95, 0.95, 0.87, 0.95, and 0.95, respectively.

ABSTRACT

Analisis teks telah menjadi salah satu proses yang penting dalam pemprosesan Bahasa tabii kerana peningkatan pesat jumlah teks berbentuk digital. Teks digital telah dijana secara berterusan dalam pelbagai format, termasuk siaran dan sembang media sosial, e-mel, artikel dan berita. Pengendalian teks ini memerlukan prosedur yang cekap dan berkesan yang mampu menangani cabaran linguistik yang timbul daripada kerumitan bahasa tabii. Dalam beberapa tahun kebelakangan ini, terdapat banyak penyelidikan untuk menggunakan ciri semantik daripada sumber leksikal bagi menangani kesukaran sinonim dan kekaburan dalam perlombongan teks seperti pengelompokan dan pengelasan dokumen. Cabaran utama mengeksploitasi sumber pengetahuan leksikal termasuk WordNet dalam cara menggabungkan pelbagai jenis hubungan semantik untuk mengenal pasti lebih banyak pewakilan semantik dan cara mengendalikan dimensi tinggi pendekatan perwakilan semantik semasa. Penyelidikan ini mencadangkan pendekatan perwakilan semantik berasaskan pengetahuan baharu bagi pengukuran keterkaitan semantik. Kaedah berasaskan pemberat untuk menggabungkan hubungan semantik dalam sumber leksikal dicadangkan dengan membentuk vektor perwakilan perkataan. Pendekatan yang dicadangkan bergantung kepada parameter topologi (kedalaman, ketumpatan, keturunan, dan salasilah) dalam taksonomi semantik. Untuk mengendalikan isu dimensi tinggi dalam kaedah berasaskan pemberat, teknik berasaskan topik baharu diperkenalkan bagi mewakili semantik perkataan dari segi topik dan bukannya konsep. Pendekatan yang dicadangkan ini bergantung pada ciri semantik dalam sumber leksikal (seperti WordNet) untuk mengendalikan isu sinonim dan kesamaran. Pendekatan yang dicadangkan dinilai dengan pengukuran keterkaitan semantik menggunakan enam set ujian standard emas. Keputusan penilaian dari segi ukuran korelasi menunjukkan bahawa kaedah berasaskan pemberat adalah lebih berkesan daripada kaedah berasaskan ciri terkini. Agar ukuran harmonik sampel tepat, nilai r dan p yang paling anomali dikira dengan menggunakan ukuran min bagi setiap set data, kaedah r dan p yang dicadangkan diuji ke atas set data MC30, RG65, WordRel353, MT287, MEN3000, dan Rgnew65 yang mana nilai r 0.82, 0.86, 0.52, 0.53, 0.89, dan 0.47; dan juga untuk p 0.80, 0.82, 0.52, 0.47, 0.82, dan 0.45 masing-masing. Keputusan pengukuran yang ditunjukkan daripada set data adalah ukuran Min standard, oleh yang demikian keputusan pengukuran pendekatan yang dicadangkan ialah 0.81, 0.84, 0.46, 0.49, 0.52, dan 0.86 untuk MC30, RG65, WordRel353, MT287, MEN30000, dan Rgnew65, masing-masing. Non-Zero (NZ) digunakan untuk menilai pendekatan yang dicadangkan bagi memastikan peratusan penjodoh kata dengan nilai perkaitan semantik yang lebih besar daripada sifar. Menggunakan data yang sama MC30, RG65, WordRel35, MT287, MEN3000, dan Rgnew65, dapatan NZ yang diperolehi dalam eksperimen ini ialah 0.96, 0.95, 0.95, 0.87, 0.95, dan 0.95, masing-masing.

TABLE OF CONTENT

REFERENCES

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2019). *A study on similarity and relatedness using distributional and wordnet-based approaches.* Paper presented at the Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Al-Mubaid, H., & Nguyen, H. A. (2006). *A cluster-based approach for semantic similarity in the biomedical domain.* Paper presented at the Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE.

AlAgha, I., & Nafee, R. (2016). Investigating the efficiency of WordNet as background knowledge for document clustering. *Journal of Engineering Research and Technology, 2*(2).

Alsmadi, I., & Hoon, G. K. (2019). Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications, 31*(8), 3819-3831.

Aouicha, M. B., Hadj Taieb, M. A., & Hamadou, A. B. (2016). Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. *Applied Intelligence*, 1-37. doi:10.1007/s10489-015-0755-x

Aouicha, M. B., Taieb, M. A. H., & Ezzeddine, M. (2016). Derivation of "is a" taxonomy from Wikipedia Category Graph. *Engineering Applications of Artificial Intelligence, 50*, 265-286.

Aouicha, M. B., Taieb, M. A. H., & Hamadou, A. B. (2016). Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. *Applied Intelligence, 45*(2), 475-511.

Aouicha, M. B., Taieb, M. A. H., & Hamadou, A. B. (2018). SISR: System for integrating semantic relatedness and similarity measures. *Soft Computing, 22*(6), 1855-1879.

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems, 165*, 346-359.

Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., & Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. *Computer Speech & Language, 56*, 107-129.

Arshad, H., Jantan, A., Hoon, G. K., & Abiodun, I. O. (2020). Formal knowledge model for online social network forensics. *Computers & Security, 89*, 101675.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. Journal of artificial intelligence research, 49, 1-47.

Bablani, A., Edla, D. R., & Dodia, S. (2018). Classification of EEG data using k-nearest neighbor approach for concealed information test. *Procedia computer science, 143*, 242-249.

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences, 32*(2), 225-231.

Batet, M., Harispe, S., Ranwez, S., Sánchez, D., & Ranwez, V. (2014). An information theoretic approach to improve semantic similarity assessments across multiple ontologies. *Information Sciences, 283*, 197-210.

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics, 44*(1), 118-125.

Bender, E. M., & Lascarides, A. (2019). Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics. *Synthesis Lectures on Human Language Technologies, 12*(3), 1-268.

Blukis, V., Paxton, C., Fox, D., Garg, A., & Artzi, Y. (2022). *A persistent spatial semantic representation for high-level natural language instruction execution.* Paper presented at the Conference on Robot Learning.

Cai, Y., Zhang, Q., Lu, W., & Che, X. (2018). A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet. *Journal of intelligent information systems, 51*(1), 23-47.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine, 9*(2), 48-57.

Carmona, C. J., Ruiz-Rodado, V., del Jesús, M. J., Weber, A., Grootveld, M., González, P., & Elizondo, D. (2015). A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans. *Information Sciences, 298*, 180-197.

Chang, C.-H., Hwang, S.-Y., & Wu, M.-L. (2021). Learning bilingual sentiment lexicon for online reviews. *Electronic Commerce Research and Applications, 47*, 101037.

Chen, G.-B., & Kao, H.-Y. (2017). Word co-occurrence augmented topic model in short text. *Intelligent Data Analysis, 21*(S1), S55-S70.

Chen, Y., Wang, J., Li, P., & Guo, P. (2019). Single document keyword extraction via quantifying higher-order structural features of word co-occurrence graph. *Computer Speech & Language, 57*, 98-107.

Choi, Y., Nguyen, M. D., & Kerr Jr, T. N. (2021). Syntactic and semantic information extraction from NPP procedures utilizing natural language processing integrated with rules. *Nuclear Engineering and Technology, 53*(3), 866-878.

Chowdhary, K. (2020). Natural language processing. In *Fundamentals of Artificial Intelligence* (pp. 603-649): Springer.

Cifariello, P., Ferragina, P., & Ponza, M. (2019). Wiser: A semantic approach for expert finding in academia based on entity linking. *Information Systems, 82*, 1-16.

Cross, V., Yu, X., & Hu, X. (2013). Unifying ontological similarity measures: A theoretical and empirical investigation. *International Journal of Approximate Reasoning, 54*(7), 861-875.

Damani, O. P. (2013). Improving pointwise mutual information (pmi) by incorporating significant co-occurrence. *arXiv preprint arXiv:1307.0596*.

Deguchi, T., & Ishii, N. (2021). *Document Similarity by Word Clustering with Semantic Distance.* Paper presented at the International Conference on Hybrid Artificial Intelligence Systems.

Dimitrova, T., & Stefanova, V. (2019). *On Hidden Semantic Relations between Nouns in WordNet.* Paper presented at the Proceedings of the 10th Global Wordnet Conference.

Du Nguyen, H., Tran, K. P., Zeng, X., Koehl, L., & Tartare, G. (2020). An Improved Ensemble Machine Learning Algorithm for Wearable Sensor Data Based Human Activity Recognition. In *Reliability and Statistical Computing* (pp. 207-228): Springer.

Du, Y., Yi, Y., Li, X., Chen, X., Fan, Y., & Su, F. (2020). Extracting and tracking hot topics of micro-blogs based on improved latent Dirichlet allocation. *Engineering Applications of Artificial Intelligence, 87*, 103279.

Echeverría, C. I. (2022). On the inertia of linguistic ideas: Revisiting the dichotomy between closed and open classes. *Language Sciences, 89*, 101445.

Escalante, H. J., García-Limón, M. A., Morales-Reyes, A., Graff, M., Montes-y-Gómez, M., Morales, E. F., & Martínez-Carranza, J. (2015). Term-weighting learning via genetic programming for text classification. *Knowledge-Based Systems, 83*, 176-189.

Esmin, A. A., Coelho, R. A., & Matwin, S. (2015). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. *Artificial Intelligence Review, 44*(1), 23-45.

Etaiwi, W., & Awajan, A. (2020). Graph-based Arabic text semantic representation. *Information Processing & Management, 57*(3), 102183.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001, April). Placing search in context: The concept revisited. In Proceedings of the 10th international conference on World Wide Web (pp. 406-414).

Faruqui, M., & Dyer, C. (2014). *Improving vector space word representations using multilingual correlation.* Paper presented at the Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.

Fellbaum, C. (1998). Wordnet: An electronic lexical database (isbn: 0-262-06197-x). In: MIT Press.

Feng, G., Li, S., Sun, T., & Zhang, B. (2018). A probabilistic model derived term weighting scheme for text classification. *Pattern Recognition Letters, 110*, 23-29.

Fischbach, J., Junker, M., Vogelsang, A., & Freudenstein, D. (2019). *Automated generation of test models from semi-structured requirements.* Paper presented at the 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW).

Flati, T., Vannella, D., Pasini, T., & Navigli, R. (2016). MultiWiBi: The multilingual Wikipedia bitaxonomy project. *Artificial intelligence, 241*, 66-102.

Fodeh, S., Punch, B., & Tan, P.-N. (2017). On ontology-driven document clustering using core semantic features. *Knowledge and information systems, 28*(2), 395-421.

Gabrilovich, E., & Markovitch, S. (2007). *Computing semantic relatedness using wikipedia-based explicit semantic analysis.* Paper presented at the IJcAI.

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). *Analysis and evaluation of unstructured data: text mining versus natural language processing.* Paper presented at the Application of Information and Communication Technologies (AICT), 2011 5th International Conference on.

Goikoetxea, J., Soroa, A., & Agirre, E. (2018). Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems, 150*, 218-230.

Guo, W., Liang, L., & Deng, T. (2017). Topic mining for call centers based on A-LDA and distributed computing. *Concurrency and Computation: Practice and Experience, 29*(3), e3776.

Gupta, V., & Lehal, G. S. (2017). *Punjabi language stemmer for nouns and proper names.* Paper presented at the Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP).

Gupta, V., & Lehal, G. S. (2013). A survey of common stemming techniques and existing stemmers for indian languages. *Journal of Emerging Technologies in Web Intelligence, 5*(2), 157-161.

Gutiérrez-Batista, K., Vila, M.-A., & Martin-Bautista, M. J. (2021). Building a fuzzy sentiment dimension for multidimensional analysis in social networks. *Applied Soft Computing, 108*, 107390.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of biomedical informatics, 48*, 38-53.

Harrow, I., Balakrishnan, R., Jimenez-Ruiz, E., Jupp, S., Lomax, J., Reed, J., . . . Wilson, J. (2019). Ontology mapping for semantically enabled applications. *Drug discovery today, 24*(10), 2068-2075.

Hatami, N., Gavet, Y., & Debayle, J. (2019). Bag of recurrence patterns representation for time-series classification. *Pattern Analysis and Applications, 22*(3), 877-887.

Hasan, A. M., & Zakaria, L. Q. (2016). QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING. *Journal of Theoretical & Applied Information Technology, 87*(2).

Hassan, S., & Mihalcea, R. (2011). *Semantic Relatedness Using Salient Semantic Analysis.* Paper presented at the Proceedings of AAAI 2011 (25th AAAI Conference on Artificial Intelligence), San Francisco.

Hong, K.-J., & Kim, H.-J. (2016). *A semantic search technique with Wikipedia-based text representation model.* Paper presented at the 2016 International Conference on Big Data and Smart Computing (BigComp).

Hotho, A., Staab, S., & Stumme, G. (2018, November 19 - 22, 2003). *Ontologies improve text document clustering.* Paper presented at the Third IEEE International Conference on Data Mining, 2003. ICDM 2003. , Melbourne, Florida, USA.

Hughes, T., & Ramage, D. (2007). *Lexical semantic relatedness with random graph walks.* Paper presented at the Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).

Jabeen, J., Abbasi, M. A., & Minhas, N. M. (2018). A Comparative Analysis of Formal languages Based upon Various Parameters. *International Journal of Engineering Research & Technology*.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169-15211.

Jiang, X., Wu, F., Li, X., Zhao, Z., Lu, W., Tang, S., & Zhuang, Y. (2015). *Deep compositional cross-modal learning to rank via local-global alignment.* Paper presented at the Proceedings of the 23rd ACM international conference on Multimedia.

Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Information Processing & Management, 51*(3), 215-234.

Jivani, A. G. (2018). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl, 2*(6), 1930-1938.

Karve, S., Shende, V., & Hople, S. (2019). Semantic Relatedness Measurement from Wikipedia and WordNet Using Modified Normalized Google Distance. In *Data Analytics and Learning* (pp. 143-154): Springer.

Kelly, M. A., Ghafurian, M., West, R. L., & Reitter, D. (2020). Indirect associations in learning semantic and syntactic lexical relationships. *Journal of Memory and Language, 115*, 104153.

Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications, 152*, 113401.

Koroleva, A., Kamath, S., & Paroubek, P. (2019). Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *Journal of Biomedical Informatics: X, 4*, 100058.

Khan, H. U., & Daud, A. (2017). Using machine learning techniques for subjectivity analysis based on lexical and non-lexical features. *International Arab Journal of Information Technology, 14*(4).

Lastra-Díaz, J. J., & García-Serrano, A. (2015). A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence, 46*, 140-153.

Lastra-Díaz, J. J., Goikoetxea, J., Taieb, M. A. H., García-Serrano, A., Aouicha, M. B., & Agirre, E. (2019a). Reproducibility dataset for a large experimental survey on word embeddings and ontology-based methods for word similarity. *Data in brief, 26*, 104432.

Lastra-Díaz, J. J., Goikoetxea, J., Taieb, M. A. H., García-Serrano, A., Aouicha, M. B., & Agirre, E. (2019b). A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence, 85*, 645-665.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database, 49*(2), 265-283.

Lee, R. S. (2020). Natural language processing. In *Artificial Intelligence in Daily Life* (pp. 157-192): Springer.

Li, P., Mao, K., Xu, Y., Li, Q., & Zhang, J. (2020). Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. *Knowledge-Based Systems, 193*, 105436.

Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering, 15*(4), 871-882.

Liang, M., Du, J., Yang, C., Xue, Z., Li, H., Kou, F., & Geng, Y. (2019). Cross-Media Semantic Correlation Learning Based on Deep Hash Network and Semantic Expansion for Social Network Cross-Media Search. *IEEE Transactions on Neural Networks and Learning Systems*.

Lin, D. (1998). *An information-theoretic definition of similarity.* Paper presented at the Icml.

Liu, J., Yang, Y., & He, H. (2020). Multi-level semantic representation enhancement network for relationship extraction. *Neurocomputing, 403*, 282-293.

Liu, X.-Y., Zhou, Y.-M., & Zheng, R.-S. (2007). *Measuring semantic similarity in WordNet.* Paper presented at the Machine Learning and Cybernetics, 2007 International Conference on.

Liu, Y., Rong, W., & Xiong, Z. (2018). *Improved text matching by enhancing mutual information.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Lu, W., Zhang, Y., Wang, S., Huang, H., Liu, Q., & Luo, S. (2020). Concept representation by learning explicit and implicit concept couplings. *IEEE Intelligent Systems, 36*(1), 6-15.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. Language and cognitive processes, 6(1), 1-28.

Macêdo, J. B., das Chagas Moura, M., Aichele, D., & Lins, I. D. (2022). Identification of risk features using text mining and BERT-based models: Application to an oil refinery. *Process Safety and Environmental Protection, 158*, 382-399.

Madani, Y., Erritali, M., & Bengourram, J. (2019). Sentiment analysis using semantic similarity and Hadoop MapReduce. *Knowledge and information systems, 59*(2), 413-436.

Madl, T., Franklin, S., Chen, K., Trappl, R., & Montaldi, D. (2016). Exploring the structure of spatial representations. *PloS one, 11*(6), e0157343.

Meng, L., Gu, J., & Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet. *International Journal of Grid and Distributed Computing, 5*(3), 81-94.

Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology, 6*(1), 1-12.

Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2022). HAKE: an Unsupervised Approach to Automatic Keyphrase Extraction for Multiple Domains. *Cognitive Computation*, 1-23.

Mohamed, M., & Oussalah, M. (2019). SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management, 56*(4), 1356-1372.

Moharir, M., & Maiya, P. (2020). Text Mining in Bioinformatics. In *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications* (pp. 63-74): Springer.

Mohd, M., Jan, R., & Shah, M. (2020). Text document summarization using word embedding. *Expert Systems with Applications, 143*, 112958.

Mosquera, R., Odunowo, M., McNamara, T., Guo, X., & Petrie, R. (2020). The economic effects of Facebook. *Experimental Economics, 23*(2), 575-602.

Mulunda, C. K., Wagacha, P. W., & Muchemi, L. (2018). *Review of trends in topic modeling techniques, tools, inference algorithms and applications.* Paper presented at the 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI).

Nag, K., & Pal, N. R. (2016). A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification. *IEEE transactions on cybernetics, 46*(2), 499-510.

Nasir, J. A., Varlamis, I., Karim, A., & Tsatsaronis, G. (2013). Semantic smoothing for text clustering. *Knowledge-Based Systems, 54*, 216-229.

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems, 182*, 104842.

Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*.

Oueslati, O., Khalil, A. I. S., & Ounelli, H. (2018). Sentiment analysis for helpful reviews prediction. *International Journal, 7*(3).

Qin, Z., Cong, Y., & Wan, T. (2016). Topic modeling of Chinese language beyond a bag-of-words. *Computer Speech & Language, 40*, 60-78.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10), 627-633.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics, 19*(1), 17-30.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10), 627-633.

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011, March). A word at a time: computing word relatedness using temporal semantic analysis. In Proceedings of the 20th international conference on World wide web (pp. 337-346).

Rago, A., Marcos, C., & Diaz-Pace, J. A. (2018). Using semantic roles to improve text classification in the requirements domain. *Language resources and evaluation, 52*(3), 801-837.

Rani, R., & Lobiyal, D. K. (2021). A weighted word embedding based approach for extractive text summarization. *Expert Systems with Applications, 186*, 115867.

Rani, P. S., Suresh, R. M., & Sethukarasi, R. (2019). Multi-level semantic annotation and unified data integration using semantic web ontology in big data processing. Cluster Computing, 22(5), 10401-10413.

Resnik, P. (1995). *Using information content to evaluate semantic similarity in a taxonomy.* Paper presented at the In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.

Ruas, T., Grosky, W., & Aizawa, A. (2019). Multi-sense embeddings through a word sense disambiguation process. *Expert Systems with Applications, 136*, 288-303.

Saif, A., Ab Aziz, M. J., & Omar, N. (2013). Measuring the compositionality of Arabic multiword expressions. In *Soft computing applications and intelligent systems* (pp. 245-256): Springer.

Saif, A., Ab Aziz, M. J., & Omar, N. (2014). Evaluating knowledge-based semantic measures on Arabic. *International Journal on Communications Antenna and Propagation, 4*(5), 180-194.

Saif, A., Ab Aziz, M. J., & Omar, N. (2015). Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features. *Natural Language Engineering, FirstView*, 1-39. doi:doi:10.1017/S1351324915000376

Saif, A., Ab Aziz, M. J., & Omar, N. (2016). Reducing explicit semantic representation vectors using Latent Dirichlet Allocation. *Knowledge-Based Systems, 100*, 145-159.

Saif, A., Ab Aziz, M. J., & Omar, N. (2017). Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features. *Natural Language Engineering, 23*(1), 53-91. doi:doi:10.1017/S1351324915000376.

Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review, 48*(2), 157-217.

Saif, A., Zainodin, U. Z., Omar, N., & Ghareb, A. S. (2018). Weighting-based semantic similarity measure based on topological parameters in semantic taxonomy. *Natural Language Engineering, 24*(6), 861-886.

Sánchez, D., & Batet, M. (2012). A new model to compute the information content of concepts from taxonomic knowledge. *International Journal on Semantic Web and Information Systems (IJSWIS), 8*(2), 34-50.

Sánchez, D., & Batet, M. (2013). A semantic similarity method based on information content exploiting multiple ontologies. *Expert Systems with Applications, 40*(4), 1393-1399.

Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems, 24*(2), 297-303.

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., . . . Quattrociocchi, W. (2017). Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences, 114*(12), 3035-3039.

Sebti, A., & Barfroush, A. A. (2008). *A new word sense similarity measure in WordNet.* Paper presented at the 2008 International Multiconference on Computer Science and Information Technology.

Sharma, A., & Jain, S. (2021). Multilingual Semantic Representation of Smart Connected World Data. In *Smart Connected World* (pp. 125-138): Springer.

Song, Y., Lin, Q., Kwon, K. H., Choy, C. H., & Xu, R. (2022). Contagion of offensive speech online: An interactional analysis of political swearing. *Computers in Human Behavior, 127*, 107046.

Taieb, M. A., Ben Aouicha, M., & Ben Hamadou, A. (2013). Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems, 50*, 260-278.

Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2013). Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems, 50*, 260-278.

Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2014). Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence, 36*, 238-261.

Taieb, M. A. H., Aouicha, M. B., Tmar, M., & Hamadou, A. B. (2011). *New information content metric and nominalization relation for a new wordnet-based method to measure the semantic relatedness.* Paper presented at the Cybernetic Intelligent Systems (CIS), 2011 IEEE 10th International Conference on.

Taieb, M. A. H., Zesch, T., & Aouicha, M. B. (2020). A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review, 53*(6), 4407-4448.

Thomas, B. A., Idris, N., Al-Hnaiyyan, A., Binti Mahmud, R., Abdelaziz, A., Khan, S., & Chang, V. (2016). Towards Knowledge Modeling and Manipulation Technologies: A Survey. *International Journal of Information Management*.

Tran, V.-K., & Nguyen, L.-M. (2021). Variational model for low-resource natural language generation in spoken dialogue systems. *Computer Speech & Language, 65*, 101120.

Utsumi, A. (2015). A complex network approach to distributional semantic models. *PloS one, 10*(8), e0136277.

Van de Cruys, T. (2018). *Two multivariate generalizations of pointwise mutual information.* Paper presented at the Proceedings of the Workshop on Distributional Semantics and Compositionality.

Verma, T., Renu, R., & Gaur, D. (2017). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems, 7*(2), 16-18.

Vijaymeena, M., & Kavitha, K. (2016). Clustering with Semantic Similarity for Text Mining. *Middle-East Journal of Scientific Research, 24*, 30-36.

Wang, R., Chen, G., & Sui, X. (2018). Multi label text classification method based on co-occurrence latent semantic vector space. *Procedia computer science, 131*, 756-764.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., . . . Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics, 87,* 12-20.

Wang, Y., Wang, M., & Fujita, H. (2019). Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems,* 105030.

Wang, Y., Wang, M., & Fujita, H. (2020). Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge-Based Systems, 190,* 105030.

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42*(4), 2264-2275.

Wei, W., & Guo, C. (2019). A text semantic topic discovery method based on the conditional co-occurrence degree. *Neurocomputing, 368,* 11-24.

Wu, Z., & Palmer, M. (1994). *V Verbs semantics and lexical selection.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Xu, Z., Harzallah, M., Guillet, F., & Ichise, R. (2019). Modular Ontology Learning with Topic Modelling over Core Ontology. *Procedia computer science, 159,* 562-571.

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation, 20*(4), 606-626.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PloS one, 15*(9), e0239441.

Yang, J., Li, Y., Gao, C., & Zhang, Y. (2021). Measuring the short text similarity based on semantic and syntactic information. *Future Generation Computer Systems, 114,* 169-180.

Yang, L., Cormican, K., & Yu, M. (2020). Ontology Learning for Systems Engineering Body of Knowledge. *IEEE Transactions on Industrial Informatics*.

Yue, L., Zuo, W., Peng, T., Wang, Y., & Han, X. (2015). A fuzzy document clustering approach based on domain-specified ontology. *Data & Knowledge Engineering, 100,* 148-166.

Zhang, C., Xie, G., Liu, N., Hu, X., Shen, Y., & Shen, X. (2021). *Automatic Hypernym-Hyponym Relation Extraction With WordNet Projection.* Paper presented at the 2021 7th International Conference on Systems and Informatics (ICSAI).

Zhang, X., Sun, S., & Zhang, K. (2018). An information content-based approach for measuring concept semantic similarity in wordnet. *Wireless Personal Communications, 103*(1), 117-132.

Zhang, X., Yang, J., Wang, R., & Li, P. (2020). A neuroimaging study of semantic representation in first and second languages. *Language, Cognition and Neuroscience*, 1-16.

Zheng, C. T., Liu, C., & San Wong, H. (2018). Corpus-based topic diffusion for short text clustering. *Neurocomputing, 275*, 2444-2458.

Zheng, Y., Li, Y., Satija, A., Pan, A., Sotos-Prieto, M., Rimm, E., . . . Hu, F. B. (2019). Association of changes in red meat consumption with total and cause specific mortality among US women and men: two prospective cohort studies. *bmj, 365*.

Zhou, X., Liang, X., Zhang, H., & Ma, Y. (2015). Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Transactions on knowledge and data engineering, 28*(2), 411-424.

Zhu, G., & Iglesias, C. A. (2018). Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications, 101*, 8-24.

Zhu, X., Yang, X., Huang, Y., Guo, Q., & Zhang, B. (2020). Measuring similarity and relatedness using multiple semantic relations in WordNet. *Knowledge and information systems, 62*(4), 1539-1569.