**Deep Learning Based Technique for Plagiarism Detection**

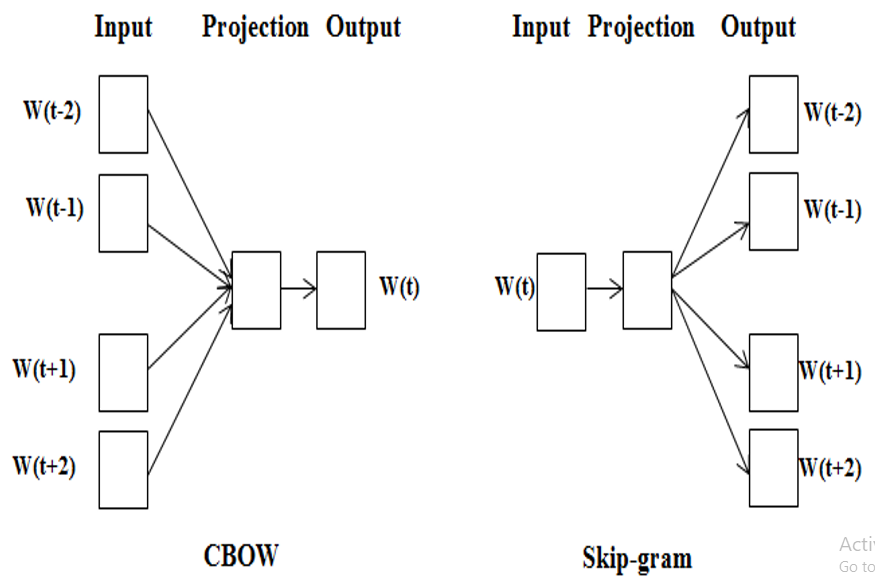
**ABSTRACT:**

Plagiarism detection is very important especially for academician, researchers and students. Although, there are many plagiarism detection tools, it is still challenging task because of huge amount of online documents. In this research, we propose to use word2vec model to detect the semantic similarity between words in Arabic language which can help in detecting plagiarism. Word2vec is a deep learning technique that is used to represent words as features of vectors with high precision. The quality of vectors representation depends on the quality of corpus used in training phase. In this paper, we used OSAC corpus for training word2vec model. Moreover cosine similarity measure is used to compute the similarity between words’ vectors. The similarity measures show how simple changes in text such as changing one word, or changing the position of verbs and nouns results with similarity value equal to 99% which provide the possibility to detect plagiarism even if the test is altered by replacing words by their synonyms or changing the words order.

.

|  |  |
| --- | --- |
| **EXISTING SYSTEM** | **PROPOSED SYSTEM** |
| **EXISTING CONCEPT:-**  Contextual word representation is very important for many Natural Language Processing (NLP) applications such as text classification, automatic summarization, information retrieval, query suggestions and plagiarism. Its importance related to the fact that it facilitates the process of finding relationships between two terms and computing their similarities. | **PROPOSED CONCEPT:** -  Basically trains a model based on a neighborhood window of words in a corpus and then projects the result onto [an arbitrary number of] n dimensions where each word is a vector in the n dimensional space. Then the words can be compared using the cosine similarity of their vectors |
| **EXISTING TECHNIQUE:-**   * **contextual representation** | **PROPOSED TECHNIQUE:**-   * **Continuous Bag-of Words model (CBOW)** |
| **TECHNIQUE DEFINITION:-**  In order to compute the contextual representation of the words, word2vec method was used. Word2vec model is a deep learning technique that is used to compute the vector representation of words using neural network with one linear hidden layer on large dataset. In addition, word2vec train the model based on sliding window, the neighbor’s words within the window are taken into consideration to compute the probability of words occurrence, and moreover the window keeps sliding over the whole corpus recursively. | **TECHNIQUE DEFINITION:-**  Continuous Bag-of Words model (CBOW), which predicts the current word depending on neighboring words and continuous Skip-gram model (Skip-gram) where the current word is the input and the model predicts the surrounding words. In both models, sliding window is used. |

**SYSTEM ARCHITECTURE**



**MINIMUMSYSTEM REQUIREMENTS**

**HARDWARE REQUIREMENTS**

* PROCESSOR : DUAL CORE 2 DUO.
* RAM : 2GB DD RAM
* HARD DISK : 250 GB

**SOFTWARE REQUIREMENTS**

* FRONT END : PYTHON
* OPERATING SYSTEM : WINDOWS 7
* IDE : Spyder3

**Conclusion:**

Plagiarism detection is one of serious tasks that represent a challenge for researchers, in this research we proposed to use word2vec model. However, Word2vec is a deep learning technique that uses large corpus for training, the output of this model is words that are represented as n dimensional vectors. Moreover, the cosine similarity between the vectors was used to detect plagiarism. In this case the similarity between vectors is contextual similarity since it depends on probability of occurrence of words within certain context. In addition to the fact that the quality of corpus determines the precision of vector representation which in turn affect the precision of plagiarism, in our experiments we used OSAC corpus. Therefore, our proposed technique is able to detect similarity between text if the changes are limited to single words replacement or order of verbs and nouns changed relatively. Accordingly, the experiment is able to detect plagiarism precision with 99% in this case.

**References:**

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In ICLR.

[2] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language and Computational Language Learning, 2007.

[3] T. Mikolov, W.T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. NAACL HLT 2013

[4] Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 2003.

[5] El Mahdaouy A., Gaussier E., El Alaoui S., 2016, Arabic Text Classification Based on Word and Document Embeddings, International Conference on Advanced Intelligent Systems and Informatics

[6] Bengio Y., Ducharme R., Vincent P., Jauvin C., A Neural Probabilistic Language Model, Journal of Machine Learning Research 3 (2003) 1137–1155

[7] Mikolov T., Sutskever I., Chen K., Distributed Representations of Words and Phrases and their Compositionality,

[8] T. Mikolov, V. Le Q., Sutskever I., Exploiting Similarities among Languages for Machine Translation, arXiv:1309.4168v1 [cs.CL] 17 Sep 2013

[9] Le Q., Mikolov T., Distributed Representations of Sentences and Documents, Proceedings of the 31 st International Conference on MachineLearning, Beijing, China, 2014. JMLR:W&CP volume 32

[10] Lin C., Ammar W., Levin C., Unsupervised POS Induction withWord Embeddings, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1311– 1316,Denver, Colorado, May 31 – June 5, 2015.

[11] Zirikly A., Diab M., Named Entity Recognition for Arabic Social Media, Proceedings of NAACL-HLT 2015, pages 176–185,Denver, Colorado, May 31 – June 5, 2015

[12] Soricut R., Och F., Unsupervised Morphology Induction UsingWord Embeddings, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1627– 1637, Denver, Colorado, May 31 – June 5, 2015.