



# **Cross-language Semantic Similarity Approach**

## **For Arabic-English Text**

**A thesis submitted to the Faculty of Computers and Information - Fayoum University in fulfillment of the requirements for the Ph.D. degree in Information Systems**

**Submitted by**

**Esraa Ahmed Abdelwahab Rslan**

Information Systems Department,  
Faculty of Computer and Information, Fayoum University

**Supervised by**

**Prof. Mohammed Helmy Khafagy**

Professor at Computer Science Department  
Dean of Faculty of Computers &  
Information Fayoum University

**Dr. Rasha Mohammed Badry**

Lecturer at Information Systems  
Department, Faculty of Computers and  
Information, Fayoum University

**2022**

## List of Publications

1. **Esraa Rslan**, Mohamed H.Khafagy, Mostafa Ali, Kamran Munir, and Rasha M.Badry, "AgroSupportAnalytics: big data recommender system for agricultural farmer complaints in Egypt", International Journal of Electrical and Computer Engineering (IJECE), 2022
2. **Esraa Rslan**, Mohamed H.Khafagy, Kamran Munir and Rasha M.Badry, "English Semantic Similarity based on Map Reduce Classification for Agricultural Complaints", International Journal of Advanced Computer Science and Applications, Vol. 12, 2021

## **Acknowledgments**

First of all, I would like to thank ALLAH for giving me the life to do an essential goal in my personal development. I want also to thank Prof. Mohamed Khafagy for his extraordinary guidance and thoughtful supervision. I would like to thank Dr. Rasha M.Badry for the practical supervision and valuable encouragement. I feel proud of my work with such great professors being with an outstanding vision for novel technological ideas.

The words cannot truly express the gratitude that I owe to my mother who raised me on morality and encourages me in every moment of despair and help me to stand again. Also, I would thank my father and my brothers for supporting me all without a doubt in my abilities, especially my beloved husband Eng. Mohamed Mahgoub endured me in the most pressures moments with all his love and patience to complete my Ph.D. degree.

Thank you all .....

## **Abstract**

Text similarity is one of the most important problems in the area of text mining and information retrieval. The importance of text similarity comes from its ability to provide the most significant information from a large text by reducing the size of textual documents. Text similarity focus on extracting the most significant information from a collection of textual documents. Most text similarities require the data to be centralized, which may not be feasible in many cases due to computational and storage limitations. The huge increase of data emerging by the progress of technology and the various sources makes automatic text similarity of a large scale of data challenging.

A model is proposed that is based on the Latent Semantic Analysis (LSA) approach to find the semantic similarity between two texts in the agricultural domain. This work is mainly focused on detecting the semantic similarity between farmer queries and agricultural dataset using LSA based on Term Frequency (TF) Weighting and Term Frequency-Inverse Document Frequency (TF-IDF) method. The model is applied to agricultural dataset that has complaints in different crops in Egypt and their solutions. We apply Support Vector Machine (SVM) classifier in MapReduce a framework for paralleling and distributing the work on the dataset to classify the dataset based on crop or problem category. The model is evaluated using several quality measures in terms of Recall, Precision, F-score, and accuracy. The results of our work showed that the classification algorithm highly improves the model. Our proposed model had outperformed the existing interest model with an accuracy of 87%.

