Agriculture Text Information System: Association Rules Mining to analyze Farmers' Complaints

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

In the Faculty of Science and Technology

By

Mostafa Ali Mahmoud Mohammed

DEPARTMENT OF STUDIES IN COMPUTER SCIENCE

UNIVERSITY OF MYSORE, MANASAGANGOTHRI

MYSURU – 570 006, INDIA

September, 2019

Abstract

In this thesis, new algorithmic models to classify text complaints have been proposed. Arabic Language has different morphology and context that make it difficult to analyze. Moreover, it is required to analyze Arabic text data automatically by using machine learning techniques. In the proposed research work, the textual data of agriculture domain is the main interest to study and analyze. Through an online portal, belonging to the Egyptian ministry of agriculture, namely VERCON, a new agriculture dataset related to farmers' complaints in Arabic text is created. The complaints' dataset in Arabic text has been collected through one online portal related to Egyptian Agriculture Research Center (ARC). The main service of ARC portal is answering farmers' queries by agriculture experts. More than 8000 complaints pertaining to forty crops have been collected.

In this thesis, statistical based and knowledge based approaches are proposed to help in extracting the features from the complaints for classification. In statistical based approach, a new weighting method called Term Class Weight-Inverse Class Frequency (TCW-ICF) is proposed. The objective of this weighting method is to enhance the classification performance of the complaints. The new method is applied on standard English datasets also as a feature selection method for evaluation. Further, an improvement is done on TCW-ICF (ImpTCW-ICF) to improve the way of selecting the key features from each class. ImpTCW-ICF is applied on the complaints' dataset and demonstrated that it outperforms other conventional methods in feature selection.

In knowledge based approach, new crop lexicon and disease lexicon are constructed. The new crop lexicon and disease lexicon are built by considering all possible terms and possible informal (slang) synonyms related to eight crops and diseases that affect the eight crops. Further, new lexicons (crop lexicon and disease lexicon) are used as agriculture domain approach for feature extraction to enhance complaints classification. Moreover, new disease ontology has been constructed based on three levels; crop name,

disease type and disease name. The ontology preserves knowledge pertaining to forty crops' diseases. Besides, the ontology facilitates discovering the relationship patterns in the complaints between the crop and its disease through mining rules. Ontology based classification approach is applied by using the new diseases' ontology. The new ontology is used for extracting the most discriminating features for every disease class.

Further, a clustering method is proposed to group unlabelled set of farmers' complaints into clusters with similar complaints. The crop lexicon mentioned earlier is used as knowledge base for feature extraction.

All the proposed methods are experimentally validated on our own newly created Arabic complaints datasets, in addition to two English standard datasets and standard Arabic dataset. A set of extensive comparative analysis with other existing contemporary models prove the superiority of the proposed methods.