STUDENT NAME: ABU AYUB ANSARI SYED

SUPERVISING PROFESSOR: DR. CHENGKAI LI

COMMITTEE MEMBER: DR. RAMEZ ELMASRI

COMMITTEE MEMBER: DR. DAVID LEVINE

DATE OF DEFENSE: JUNE 14, 2017

TIME OF DEFENSE: 10:00 AM

ROOM NUMBER: 301 ERB

TITLE OF ABSTRACT: EVALUATION OF A FACTUAL CLAIM CLASSIFIER WITH AND WITHOUT USING ENTITIES AS FEATURES

COMPLETE ABSTRACT:

Fact-checking in real-time for events such as presidential debates is a challenging task. The first and foremost task in fact-checking is to find out whether a sentence is factually check-worthy. The UTA IDIR Lab has deployed an automated fact-checking system named ClaimBuster. ClaimBuster has a core functionality of identifying check-worthy factual sentences.

Named entities are essentially an important component of any textual data. To use these named entities as a feature in a classification task, it is required to link them to labels such as person, location and organization. If we want the automated systems to read and understand the natural language like we do, the system must recognize the named entities that are mentioned in the text.

The ClaimBuster project, in classifying the sentences of the presidential debates has categorized the sentences into three types, namely check-worthy factual sentences (CFS), non-factual sentences (NFS) and unimportant factual sentences (UFS). This categorization helps us in making this supervised classification problem as a three-class problem (or a two-class problem, by merging NFS and UFS). ClaimBuster, in the process of identifying check-worthy factual claims has employed named entities as a feature along with sentiment, length, words (W) and part-of-speech(POS) tags in the classification models. In this work, I have evaluated the classification algorithms such as Naïve Bayes Classifier

(NBC), Support Vector Machine (SVM) and Random Forrest Classifier (RFC). The evaluation mainly constitutes the comparison of performance of these classifiers with and without using named entities as a feature. We have also analyzed the mistakes that the classifiers have made by comparing two sets of features at a time. Therefore, the analysis consists of 18 experiments constituting 3 classifiers, 2 classification types and 3 sets of feature comparison. We see that the presence of named entities contributes very little to the classifier, but also that their presence is subdued by presence of better performing features such as the part-of-speech (POS) tags.