**SOMIDIA – Current Thesis Comparison**

|  |  |  |
| --- | --- | --- |
|  | **SOMIDIA** | **CURRENT THESIS** |
| **RESEARCH** | | |
| **Thesis Problem** | Currently, there is no disaster management system in the country that implements an information extraction component that will allow the extraction of information for disaster management from different online sources and SMS reports. And also, no existing system is taking advantage of the information made available through social networking sites. Furthermore, no existing system in the country can provide reliable crowd-sourced information for the reason that no robust and proper data validation technique in terms of data gathering for this domain has been implemented yet. | The existing information extraction system present in SOMIDIA has limitations when it comes to handling text because it can only handle English text. The system still has problems when handling Filipino text for the reason that the information extraction done on text written in Filipino is using keyword-based searching. Currently, there are no stable tools that can handle Filipino text (Cheng et. al., 2011). Social media data in the Philippines shows inconsistency because when Filipinos communicate or post in social networking sites they use different languages such as English, Filipino, Taglish and “TXTSPK”, especially today, when people with different language and culture communicate on the same medium (social networking sites) (Ghedin, 2011). Another problem with social media data is the existence of micro-posts, or short posts, such as tweets which do not have much contextual information and tend to be less grammatical. (Maynard et. al., 2012). |
| **General Objective** | To develop a disaster management system that extracts and validates data from various sources in near real-time and provides visualization of the information to authorities concerned. | To develop an information extraction system that extracts relevant information from disaster-related texts from social media and takes into consideration the different available variations in the Filipino language. |
| **Scope and Limitations** | * An adaptive IE system is integrated, which also serves as a means of gathering data * The system will cover the whole stretch of Metro Manila * Main focus: Top 5 disasters that have affected the country in terms of frequency and affected population for each year during 2007-2010   + natural disasters (typhoons, floods, droughts, landslides, earthquake)   + human-induced disasters (armed conflict)   + ad-hoc disaster situations * Information extracted:   + place of the event   + time/duration of the event   + persons involved in the event   + type of event   + description of the event * Source of data   + local news sites   + Twitter   + Facebook   + SMS reports * The system is capable of handling English text * Filipino / Taglish / text speak text will be handled with keyword-based searching | * Disasters:   + natural disasters (typhoons, floods, droughts, landslides, earthquake)   + human-induced disasters (fire) * Information extracted:   + Type of event   + Place of the event * Source of data   + Twitter   TinTin’s Assumptions:   * The system will only handle the IE module * The collection of tweets is already here * Take advantage of the keywords listed in SOMIDIA’s paper * Add objective/SL pertaining to disaster |
| **SYSTEM** | | |
| **Overview** | * SOMIDIA is a crisis mapping system that focuses on plotting the authentic crisis events on an interactive map in near real time * Takes advantage of crowdsourcing as its means of input * Plots the extracted information from online sources * Users may submit disaster reports or request for help via the system's website or send an SMS to the system's dedicated number * Validation of reports is done via crowdsourcing -- registered users | * The system to be developed will serve as the information extraction module of SOMIDIA for Filipino disaster-related reports |
| **General Objective** | To develop a crisis mapping system that is able to provide visualization of validated crisis events based on crowd-sourced information in near real time. | To develop the information extraction module of SOMIDIA that extracts relevant information from disaster-related texts from Twitter data and takes into consideration the different available variations of the Filipino language. |
| **Specific Objectives** | 1. To automatically retrieve textual resources from online sources; 2. To extract relevant information from these textual resources; 3. To validate the gathered information; 4. To score and rank the point of origin of the gathered information; 5. To determine the weights which will be used to quantify the credibility of extracted information; 6. To plot the incident reports in an interactive map in near real time; 7. To present the interactive map and list of reports in a manner most relevant to the user; 8. To facilitate external validation of submitted reports. | 1. To classify if the tweets is disaster related or not; 2. To classify the tweets whether it is a fact or opinion; 3. To extract relevant information from the factual tweets; 4. //To evaluate the system’s performance; |
| **Approaches for Filipino IE** | Keyword-based searching | No decided approach yet. |

|  |  |
| --- | --- |
| **SOMIDIA ARCHITECTURE** | **CURRENT THESIS ARCHITECTURE** |
|  | **Information Extraction Module**  **Tweets**  **Pre-Processing Module**  Text Normalization  Text Classification  **Information Extraction Engine**  **???**  **Output**  **Output Generation** |