KEYSTONE COST Action IC1302

Short Term Scientific Mission Report

STSM Topic: Keyword-based search in non-structured data sources of relevant meteorological events

STSM Applicant: Dario Stojanovski, Faculty of Computer Science, Ss. Cyril and Methodius University, Skopje, Macedonia

stojanovski.dario@gmail.com

STSM Host: Associate Professor Jose Ramon Ríos Viqueira, University of Santiago de Compostela, Santiago de Compostela, Spain

jrr.viqueira@usc.es

Period

04.07.2016 – 17.07.2016

Purpose of the STSM

One of the main purposes of the STSM was to establish the collaboration with prof. Viqueira and prof. Alberto Bugarin Diz. Another important goal was to determine the specific topic and steps of the research project. As a result, we decided to start work on techniques that will enable matching and visualizing several different types of information related to severe weather events. The main contribution of this work is to define a method that will enable determining damage from severe weather events from social data and enable searching for areas with specific damage levels.

The practical applications of such system are multiple fold, especially if the proposed architecture is to be set up in a real-time environment. The system can be used to analyze the behavior of people during such events from their social media activity. It could be used to provide useful information in the process of discerning patterns in user behavior which could in turn, be utilized to better respond to such crisis in the future. In the real-time case, the system would be able to estimate the caused damage in specific regions which cannot always be easily determined from meteorological data alone. This could give better insight to emergency response teams on where to focus their efforts.

There is a lot of work that analyze the effect that popular events invoke on social media. A great deal of them are related to the Sandy hurricane which affected the US east coast and specifically, New York in 2012. However, many of them provide more or less manual analysis of the aggregated data. In our work, we will strive to automate the analysis in order to provide real-time valuable information during meteorological events.

Description of the work carried out during the STSM

The duration of the STSM was two weeks. In the first week, the main focus was to outline the work to be done during the stay and afterwards and also detail the specific steps to be taken. In the first week, several meetings were held. The host researchers gave presentation on the work they are doing related to the joint research. Additionally, I gave a presentation of the research I have done so far that was related to the topic of the STSM. Most of the work was orientated on obtaining relevant work to the research that was defined as the main goal of our work. Several papers were identified as relevant and we focused on studying them in detail. For this part, it was important that a lot of papers from several publishers were freely available which would have not been at my home institution.

We decided on analyzing hurricanes as they are relatively long severe weather events that can cause damage on a large scale. Especially interesting was the Sandy hurricane because it affected a densely populated area, one that is a well known source of an abundance of social data. A dataset of Twitter messages was identified to be used in the research that was collected during the hurricane. However, Twitter does not allow for sharing of tweet content, only tweet IDs. Consequently, we had to utilize the Twitter API in order to download the actual tweet content. The dataset contains over 15 million tweets, out of which less than 11 million were available for download. As Twitter limits the number of tweets that can be retrieved in a certain period of time and because of the huge number of messages, the retrieval of the dataset took about two days. At the end of the week, we had a conference meeting with my mentor, prof. Gjorgji Madjarov and decided on what direction should we take in the research project.

At the beginning of the second week, we downloaded meteorological and damage data related to Hurricane Sandy. We started working on a web page that can be used to visualize the data that was previously acquired. We showed several different visualizations of the geographical data, both separately and jointly. Additionally, we visualized data on different time scales during the hurricane. The work was done using the Google Maps API and the Django application framework. The application is able to generate heatmaps from damage data. Moreover, the same technique was used to visualize Twitter data as well. One interesting visualization offers heatmap visualization of social data on different time scales and jointly shows the path that the hurricane took along the east coastline. This can show how the distribution of social activity changes during the event. I also got familiar with the QGIS tool that enables more complex handling and visualizations of geographical data. With

the help of prof. Viqueira, I learned about several useful features of this program and how can we utilize the software to better present the data.

Moreover, in order to provide more meaningful analysis of the social data, we utilized several models for recognizing and classifying sentiment and emotions in tweets which we have already developed. The architecture of the models is based on the newly emerged deep learning techniques such as convolutional and recurrent neural networks. Additionally, we have already developed a hybrid between the two aforementioned networks. During the stay, we trained a model with a CNN. The network was trained on previously downloaded Twitter messages which are not specifically related to meteorological events. As the model is based on a deep learning technique, training took some time to complete considering the number of parameters it needs to optimize and the size of the data it needs to process. For the purpose of this research, we considered training an emotion recognition model on tweets that contained similar words to those in the dataset related to Sandy. However, the final decision on whether to use these or any random tweets is still not made. In the following period, we will try the other network architectures and decide on the best performing one.

In related work, only sentiment was included in the analysis of social data. We believe that recognizing several distinct emotions from textual data can provide more detailed insight into the way people reacted to the storm. This model is easily adaptable to a stream setting as well, where data can flow continuously through the system. However, depending on the volume of messages, deep learning models, even relatively simple ones such as those we used here, would require more resources, specifically, GPU hardware.

Description of the main results obtained

A significant result of the STSM is the foundation of a scientific and research collaboration between myself and my mentor, prof. Madjarov on one side, and prof. Viqueira and prof. Bugarin on the other side. We identified a suitable research project, detailed the tasks we need to accomplish and outlined the topic of the potential publication. We completed a detailed survey of related work and identified potential deficiencies in previous works. As a result, we were able to find what aspects can be improved and what can be added to provide more meaningful analysis.

We started work on techniques for more relevant joint visualization of meteorological, damage and social data for severe weather events. Concretely, the system was developed on top of data for hurricane Sandy. From the initial results, we observed a relation between the distribution of tweets related to Sandy and the path and intensity of the hurricane itself. We observed that a lot of tweets originated from New York, but also a significant number came from the area around Florida. This high number of tweets around Florida diminished as the storm passed through this region after which New York remained the only dominant source of social data.

Additionally, we trained a convolutional neural network on a training set of 10K samples of emotionally labeled tweets which will be used to recognize possible expressed emotions in the dataset related to Sandy.

Future collaboration with the host institution

We have considerable plans for future collaboration. The main goal is to continue the work that we have started. We have outlined several steps to be taken upon returning from the STSM. First, we plan on developing a techniques that will allow us to determine the damage level from the Twitter messages themselves. The technique will be evaluated using the actual damage data. We plan on creating geographical hexagons and color them based on the damage level, both reported by the reported damage and estimated from the social data. We will conduct detail analysis of the sentiment and emotionally labeled tweets and observe the geographical and temporal component of the emotions during the hurricane. At this time, the system is developed to work in an offline environment. However, we plan on adapting it for real-time analysis of the meteorological and social data in order to estimate areas that were severely impacted by storms and similar weather events. We will have regular meetings with the host institution.

Foreseen publications/articles resulting from the STSM

The final goal of the proposed joint work is to publish a paper at a conference or journal paper. We have continued working on the project remotely. Once we have all the results from the analysis of the damage, meteorological and social data, we will begin drafting the potential publication and start working on writing the paper.

Skopje, Macedonia, 10 August 2016 Dario Stojanovski