CHAPTER 17

Social Networks VGI: Twitter Sentiment Analysis of Social Hotspots

Dario Stojanovski*, Ivan Chorbev, Ivica Dimitrovski and Gjorgji Madjarov

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
*stojanovski.dario@gmail.com

Abstract

The enormous amount of data generated on social media provides vast quantities of geo-referenced data. Volunteered Geographic Information (VGI) originating from social networks has produced new challenges for research and has opened opportunities for a wide range of use cases. Smartphones with built-in GPS sensors enabled users to easily share their location and with the growing number of such devices available, VGI data is expanding at a rapid rate. Twitter is one of the most popular microblogging services. It's a social network that enables access to the data that is being created on the platform. It also allows for real-time retrieval of data from a given geographic area.

In this paper we give an overview of a system for detecting and identifying social hotspots from Twitter stream data and applying sentiment analysis on the data. Utilizing the Twitter Streaming Application Programming Interface (API), we collected a significant number of Tweets from New York and we evaluated the quality of the retrieved data. In this paper, we outline advantages and disadvantages of using various clustering algorithms over the data for this purpose, namely hierarchical agglomerative clustering and DBSCAN. We also elaborate on techniques for identifying social hotspots from spatially localized

How to cite this book chapter:

Stojanovski, D, Chorbev, I, Dimitrovski, I and Madjarov, G. 2016. Social Networks VGI: Twitter Sentiment Analysis of Social Hotspots. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 223–235. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.q. License: CC-BY 4.0.

clusters. Finally, we present a deep learning approach to sentiment analysis used to determine the attitude of users participating in the identified social hotspots.

Keywords

social hotspots, sentiment analysis, Twitter, VGI, visualization, geo-clustering

Introduction

Volunteered Geographic Information (VGI) as coined by Goodchild (2007) is defined as the harnessing of tools to create, assemble and disseminate geographic data provided voluntary by individuals. VGI received a lot of traction in recent years with the continuing evolution of Web technologies that provide for easier user participation in the creation of such data with users assuming a more active role in the creation of VGI data (Sui 2011). This idea stimulated the popularity of a number of services such as Wikimapia, OpenStreetMap and many others. Wikimapia is based on the same concept as Wikipedia and it contains over 23.8 million objects. Flickr on the other hand, a system for hosting images allows users to upload geo-tagged photos on the platform with the corresponding latitude and longitude pair that are associated with the picture. Alternatively, people can also act as sensors and just provide their location through the use of various social network services such as Facebook, Twitter, Foursquare etc.

Taking advantage of VGI or otherwise known as User Generated Spatial Content (UGSC) leads to a vast number of applications ranging from public health, early disaster warning and crisis management to various types of analytics useful to marketing agencies and companies.

Social networks and microblogging platforms have attracted the attention of users on a huge scale over the recent years. Platforms such as Facebook, Instagram, Foursquare and many others generate massive amounts of data. These services also allow users to geo-locate the information they share on these networks. This lead to popularization of Social Media Geographic Information (SMGI) which refers to the geo-referencing of multimedia data extracted by social media applications. With the spread of mobile devices equipped with GPS sensors, it become even more accessible for users to share their location in order to provide more context to the content they are sharing. As of 2015, Twitter, the most popular microblogging platform of all, has over 300 million monthly active users and generates over 500 million messages on a daily basis[1]. One key advantage of Twitter is its real-time component, positioning the plat-

[1] https://about.twitter.com/company

form as one of the most up to date data source, as witnessed by its ability to break news before other sources.

Bloggers in the Twitter community use the platform to express their views and ideas on different topics, share thoughts on their daily activities, celebrity gossip etc. Although only a small percentage of Tweets (1.2%) (Dredze 2013) contain exact location, the sheer volume of messages generated every day, makes Twitter a gold mine for mining VGI data. Since users Tweet about events around them in real-time, we could tap into this information stream and identify social hotspots as they are emerging. Furthermore, applying sentiment analysis on the content shared related to a social hotspot, can provide additional insight about the place or event.

Sentiment analysis on social media has wide applications as it can be used to provide feedback for the reaction products and services receive, the public opinion towards different candidates during political elections etc. The presented work, focuses on analysis of sentiment related to social hotspots.

Related Work

A lot of research has been conducted to explore the various applications of volunteered geographic data from social media (Sui 2011). Companies have also showed interest in the area along with the field of sentiment analysis because of its potential to provide valuable insight into people's reaction regarding related products and the distribution over geographical areas (Liu 2014).

Dredze et al. (2013) explored the application of Twitter geo-located data to public health. They developed a system that infers structured location information from Tweets and showed how this information can be used for influenza tracking. However, their approach only detects location on city level and it's not able to detect finer grained locations. In the work of Li (2013), he addressed the issues of extracting local information and discovering communities of interest in local social media. Bosch et al. (2013) propose a visual analytics approach to facilitate sensemaking of geo-located microblog posts by enabling analysts to create automatic methods for extracting messages and by applying those methods when monitoring topics of interest.

Twitter as a source of geo-data has been used in various domains, many of which focus on event detection from Twitter data. Abdelhaq et al. (2013) presented a framework for detection of localized events in real-time from Twitter streams and tracking their evolution over time. The proposed system uses both geo-located and non-geo-located Tweets to identify event describing words, but only geo-located Tweets are used to determine the spatial distribution of such words. Spatial and temporal characteristics of keywords are continuously extracted to identify meaningful candidates for event descriptions. The system selects words that show bursty frequency in the current time frame and have local spatial distribution. Keywords are then clustered by their spatial

signatures and clusters are scored to show how likely is that they represent a localized event. However, this approach does not perform very well in situations when there are multiple geo-terms within the same text.

Kisilevich et al. (2010) developed a new version of the DBSCAN clustering algorithm for analysis of places and events using a collection of geo-tagged photos. The assumption is that a high photo activity in a specific area is indicative of an interesting place or an ongoing event. The proposed algorithm addresses an issue that is specific to identifying social hotspots, which occurs when a significant portion of the samples in a cluster originates from a single user. In our work, we also utilize the DBSCAN clustering algorithm and tackle this issue in the cluster detection phase where features are generated indicative of the number of users in a cluster. Evaluation of the approach is done on Flickr images from Washington, D.C.

Walther et al. (2013) propose a system that detects geo-spatial events from the Twitter stream. The proposed system create clusters or EventCandidates in a manner similar to the DBSCAN clustering algorithm. Clusters are further analyzed to detect if any overlaps have occurred, both spatially and temporally. Several hand-crafted features were developed that the authors consider to be indicative of an actual event and these features are extracted from each cluster. Finally, an evaluation whether a cluster represents an actual event or not is made using a machine learning approach. Our system builds on the work of Walther et al. (2013). Additionally, we extend the system by applying sentiment analysis on the identified social hotspots.

Data Retrieval

Retrieving Twitter messages is available through the Twitter Application Program Interface (API) which offers a variety of REST endpoints. Nonetheless, in order to continuously collect Tweets from a certain geographic location, generating repeated REST calls is infeasible due to Twitter rate limits. As a result, we must utilize the Twitter Streaming API that gives low latency access to Twitter's global stream of data. The stream can return Tweets originating from a set of users or messages that contain certain keywords. However, the possibility of supplying the Streaming API with a filter specified by a set of spatial bounding boxes defined by latitude and longitude pairs is of interest in our work. This filter provides real-time access to all messages originating from a given geographic area.

A Tweet and its location are available through the Twitter public streams if the user explicitly consents to sharing the location in the post. Users can enable locations on their devices and provide exact coordinates obtained from a GPS sensor of the device they used to post the message. Additionally, Twitter enables for manual embedding of places in messages. Places on Twitter have specific IDs, defined by a bounding box and can be of several different types

(city, POI, country). For a Tweet with an embedded place to be returned, the bounding boxes of the place and the filter applied in the stream must intersect. One drawback of retrieving Tweets with places is that the user location may not be the same as the one mentioned in the Tweet. A user can post a Tweet containing a mention of a place while Tweeting from somewhere else. Furthermore, many places refer to larger areas, cities or even countries which may feed the stream with Tweets that are outside of the defined bounding box. As a result, it is necessary to additionally filter the incoming data. MongoDB and its document model is probably the most suitable database system for storing social media posts. In addition, it supports temporal and geo-spatial indexes which is essential to the task at hand as we are dealing with geographical and temporal data.

In future work, it would be valuable to explore enriching the data with Twitter messages that are not explicitly geo-tagged. As for geo-tagged Tweets, the percentage of geo-tagged ones with exact coordinates goes as high as 1.2% (Dredze 2013), while 1.3% contain a Place object. In order to increase the utilization of the geo-referenced information generated on Twitter, one must look beyond explicitly volunteered geo-data. Users often include references to geographical information in the content they post on the Web, without tying it to specific coordinates. Location recognition in social media is a challenging problem, even more so in Twitter due to the 140 character limitation and the abundance of abbreviations, informal language or terms used only in the Twitter community.

In order to enrich dataset with VGI, one must first define a set of keywords to track using the Streaming API, as the number of keywords that can be fed to the API is limited to 400. In order to get Tweets most related to social hotspots, the stream should be fed with keywords relevant to social hotspots and with words related to the area that is being monitored. Liu (2014) developed an extensive system to disambiguate and identify locations mentioned in text and for estimating user location out of their activity. The approach relies on sequential learning methods to automatically learn the relations between parts of locations where the classifiers are fed with hand-crafted features.

Dataset

The system that is showcased in the remainder of the paper is based on Twitter data from New York between February 22 and April 16 2015. We set the bounding box to the following longitude and latitude pairs: (−74, 40), (−73, 41). New York City is chosen because it's one of the most active cities Twitter-wise. Also, we assume that the majority of the messages will be in English. New York's big population and its dense social places structure pose both difficulties and advantages for mining geo-data. On one hand, the abundance of social places suggests that a relatively high number of Tweets will be related to social places.

On the other hand, the dense structure poses challenges to precise clustering of Tweets.

For the above mentioned period, we collected 4,125,542 Twitter messages, generated from a total of 226,114 distinct users. Out of these, 3,274,724 messages contained exact coordinates, while the others had a place entity only, which were attached manually. However, we observed that among these Tweets that also had place entities attached, a significant portion was outside of the defined bounding box. Upon filtering, the dataset was reduced to 2,350,739 messages.

In the set collected, we observed that place entities generally are related to greater geographical areas. For example, places such as 'Manhattan', 'New York' or 'Brooklyn' appear very often, as opposed to points of interests. Such places are insignificant to our analysis and have to be filtered out because they refer to a very broad area. Only 9,119 Tweets with embedded POIs that are within the defined bounding box were retrieved, while 269 messages have POIs outside of the bounding box.

System Overview

The presented system in this work monitors Twitter streams for a defined geographic area and identifies social hotspots as they are emerging. Figure 1

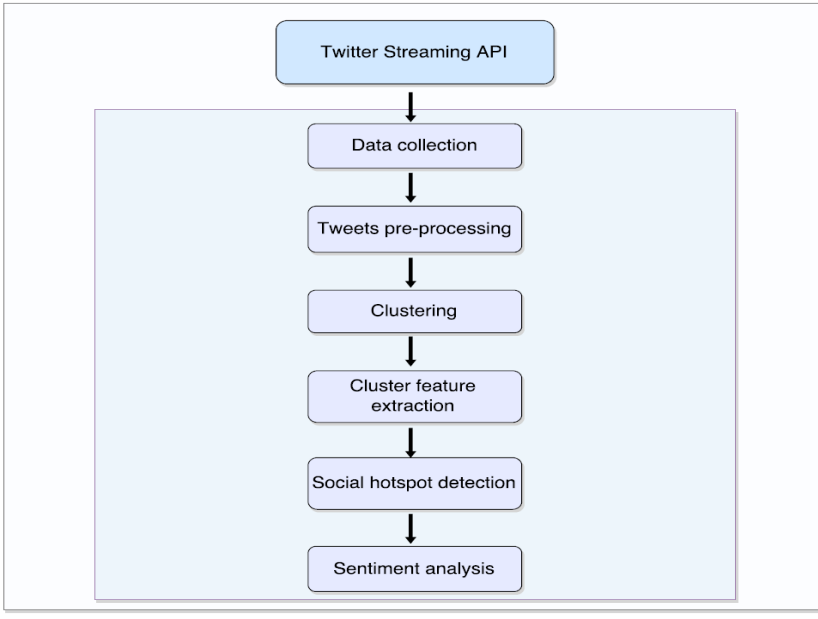

Figure 1: System architecture.

depicts an overview of the system architecture. The system can be broken down to the following key components:

- Data retrieval – connects to the Twitter Streaming API and collects messages from a certain geographic area.
- Tweet pre-processing – cleans Tweets from noisy tokens and characters.
- Cluster generation – creates clusters or social hotspots candidates from Tweets retrieved in a limited time period.
- Cluster feature extraction – extracts relevant features from Tweets in a social hotspot candidate
- Social hotspot detection – analyses clusters and evaluates using machine learning whether they represent a social hotspot or not.
- Sentiment analysis – analyses the sentiment of the Tweets in the social hotspot clusters.

Social Hotspot Detection

A social hotspot is a geographic POI which attracts the attention of many people in a limited period of time. In the context of Twitter, detecting social hotspots requires locating places with highly concentrated activity. In order to identify social hotspots from Twitter streams, clusters of geographically close Tweets must be created. There are several ways of generating such clusters, few of which are elaborated in the remainder of this work.

Clustering is an unsupervised machine learning method where samples from a given set of data are grouped based on features describing each sample. For purposes of the system described here, only the geographical component of the Tweets is taken into consideration for the clustering phase. Clustering algorithms compute distance between samples from the dataset. Since we are dealing with geo-data in relatively confined areas, the most appropriate metric is Euclidean distance.

Algorithms such as the K-means algorithm that require the number of clusters to be predefined are not appropriate for the specific problem, because the number of clusters or social hotspots candidates cannot be anticipated. One way of overcoming this issue is by using hierarchical agglomerative clustering. Each observation or Tweet starts-off as single cluster. Iteratively, pairs of clusters are merged together as the algorithm moves up the hierarchy. The result is a dendogram from which the threshold value can be observed which is used to cut the dendogram and to prevent it from building into the complete hierarchy. In order to get sufficiently localized clusters the threshold has to be set to a very small value. However, the complexity of hierarchical clustering ranges from $\Theta(N^2)$ to $\Theta(N^3)$, depending on the selected linkage criteria. Another deficiency is that it requires the pairwise distances of all the observations in the set. Computing pairwise distances has a $\Theta(N^2)$ memory complexity, which can be infeasible if the number of input data is huge.

Another appropriate clustering technique is DBSCAN. It is a density based clustering algorithm, not limited to shapes of clusters and only relies on a neighborhood count parameter (minPts) and a neighborhood distance ε. The general algorithm classifies each sample as core points, reachable points and outliers as follows:

- Core points have at least minimum number (minPts) of other points within ε distance
- A point p is density reachable if there is a point q and a path $p_1 \ldots p_n$ where $p_1 = p$ and $p_n = q$ and p_{i+1} is directly reachable from p_i and is a core point
- A point is an outlier if it is not reachable from any other point

Points with a density above the specified threshold are constructed as clusters. DBSCAN handles outliers well, and has been proven as very effective in processing very large databases. It is by far most suitable for the task at hand as it gives close control over what is considered a social hotspot candidate. ST-DBSCAN is also a density-based algorithm for clustering spatial-temporal data. Birant et al. (2007) first proposed this approach as an extension on the existing DBSCAN in relation to the identification of core and noise objects and adjacent clusters. ST-DBSCAN takes into consideration the non-spatial, spatial and temporal attributes of the data. This is especially important in this case as social hotspots are not fixed in time. The algorithm requires two additional parameters, the distance parameter for non-spatial attributes and Δ_ε which is used to prevent the discovering of combined clusters. Additional modification is that a region is dense if the minPts criteria is satisfied by both of the distance parameters. ST-DBSCAN is also efficient at handling noise points when there are clusters with different densities. In Figure 2, the figure depicting clusters generated using hierarchical clustering, only clusters containing at least minPts messages are presented. Different marker colors are used for better clarity. We observe that DBSCAN generates less clusters than using hierarchical clustering. The generated clusters need to be further analyzed in order to determine if the Twitter messages refer to an actual social hotspot or are just random non-related posts or conversations. For this, we borrow on the work of Walther et al. (2013).

They developed several features divided into two categories. We only present the ten most effective features.

- Unique posters – the total number of unique users.
- Common theme – calculates word overlap between different Tweets in the cluster.
- @ Ratio – the number of user mentions relative to the number of Tweets
- Unique coordinates – the total number of unique coordinates within a cluster
- Ratio of Foursquare posts – fraction of Tweets originating from Foursquare
- Tweets count – total number of Twitter messages

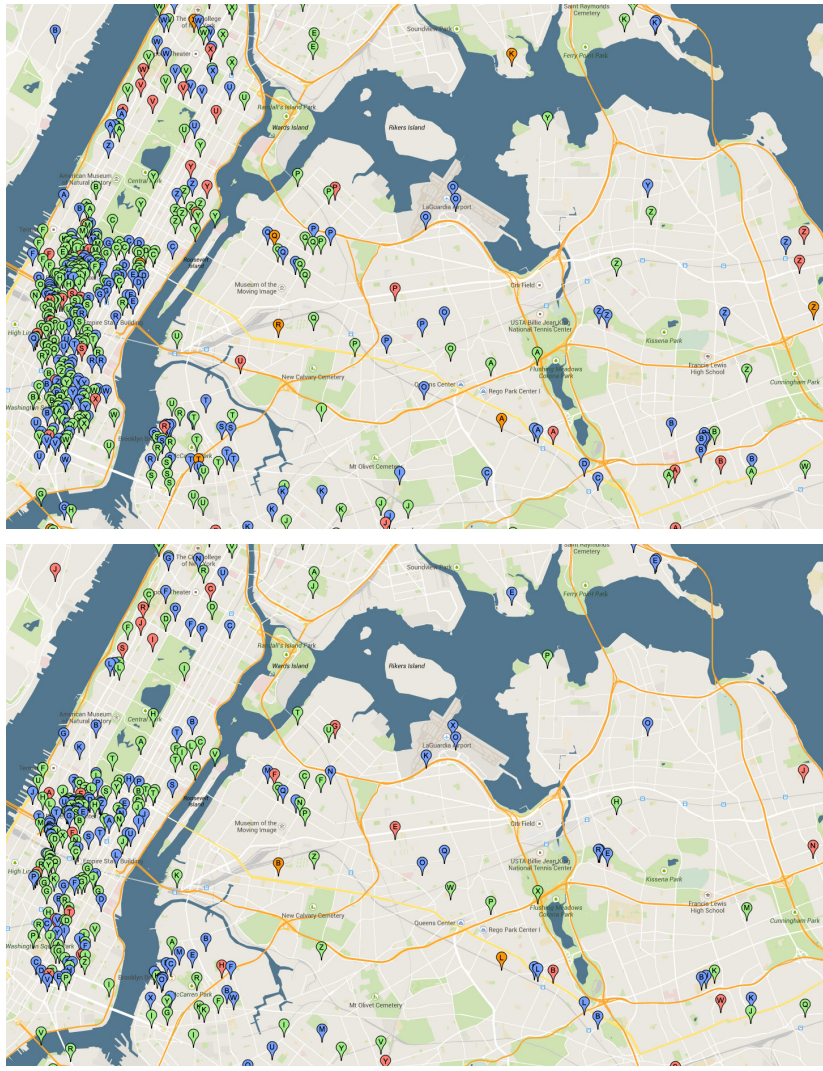

Figure 2: Clusters generated with hierarchical and DBSCAN clustering. Each marker represents a separate social hotspot candidate.

- Semantic Category – whether the cluster belongs to one of several predefined categories
- Subjectivity – indicates whether users share subjective posts or just various information
- Positive sentiment – indicates positive sentiment
- Ratio of unique posters – the number of unique users in relation to the number of Tweets

The effectiveness of the described features are experimentally evaluated in the work of Walther el al. (2013). They manually annotated 1000 clusters with binary labels in respect to whether they represent a real-world event or not. Textual features proved more significant than the Other group, but a combination of both feature categories provides for best performance. Walther et al. (2013) used three different machine learning algorithms, specifically Naive Bayes, Multilayer Perceptron and C4.5 decision trees. However, it would be beneficial to evaluate the effectiveness of Support Vector Machines and other machine learning approaches.

Sentiment Analysis

Sentiment analysis is the task of identifying human emotion in text. Social networks and media sparked interest in sentiment analysis amongst both academia and industry as users often share opinions and feeling on social networks (Pak 2010). Observing sentiment regarding social hotspots can provide valuable information about the popularity of a place or an event that is taken into consideration.

So far, Twitter sentiment analysis has heavily relied on hand-crafted features, which are both incomplete and too domain specific and depend on lexicons with sentiment polarity. Additionally, the process of manual feature generation is time-consuming and requires extensive domain knowledge. Deep learning techniques on the other hand, automate the feature generation and are more robust and flexible when applied to various domains. Convolutional Neural Networks (CNN) have been shown to achieve state-of-the-art results in sentence classification and specifically in sentiment analysis (Kim 2014), (dos Santos 2014).

We have developed an architecture for sentiment analysis that uses a CNN with multiple filters with varying window sizes. The model is built on the work of (Kim 2014) where they report state-of-the-art performances on 4 out of 7 sentence classification tasks. It consists of one convolutional layer and a max-over-time pooling layer which outputs a fixed sized vector. This vector is then fed to a three layer feed-forward network with two non-linear layers and a softmax output layer which gives the probability distribution over the sentiment classes. The architecture maps each token in a given Tweet to an appropriate word representation. The approach leverages large Twitter corpora for unsupervised learning of these word representations, which capture syntactic and semantic characteristics of words. Instead of doing the pre-training of word embeddings ourselves, we use available word vectors.[2] We continuously update word vectors by back-propagation during training time and by doing so we capture and encode sentiment information into the word embeddings. We train

[2] http://nlp.stanford.edu/projects/glove/

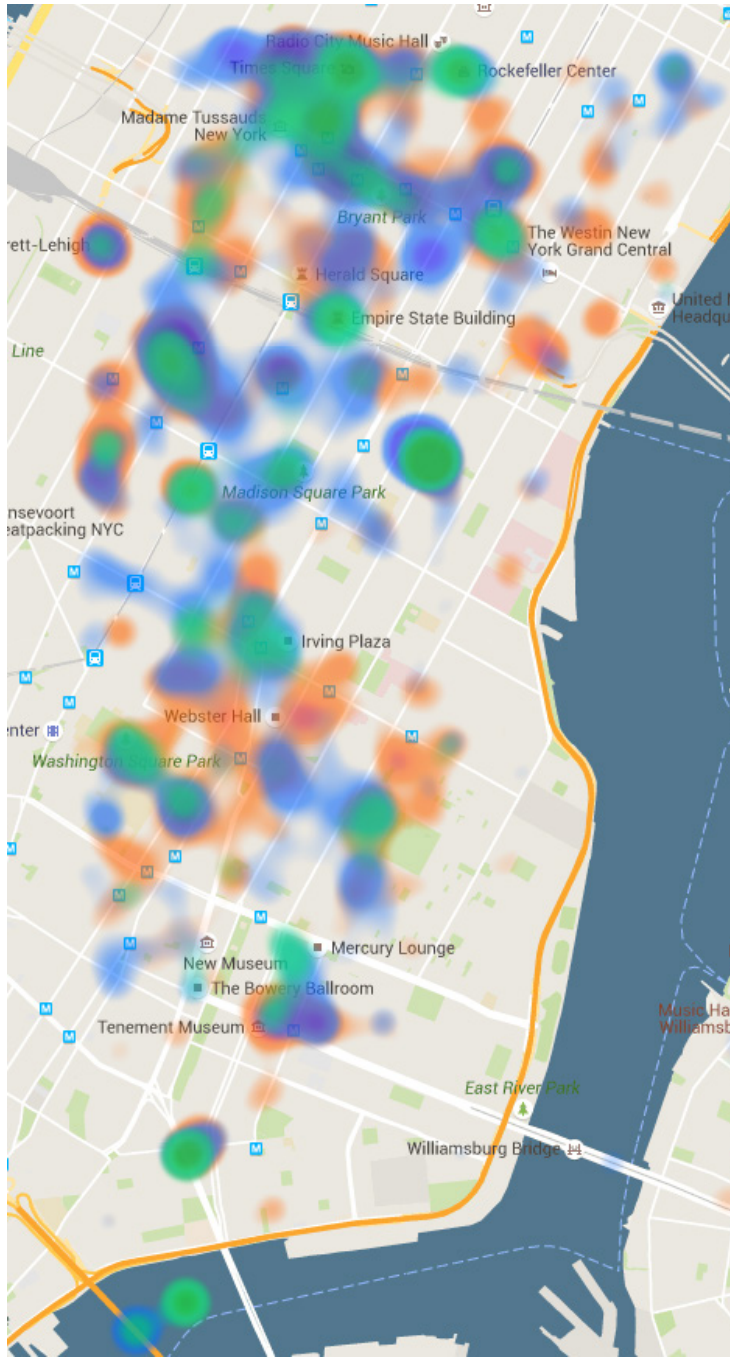

Figure 3: Sentiment heatmap.

the deep convolutional neural network with manually annotated Tweets provided by the Sentiment Analysis in Twitter task on SemEval-2015. The neural network in our system classifies Tweet sentiment into 3 classes, where the labels can be positive, negative or neutral. For future work, it would be interesting to use the proposed architecture for emotion identification which may provide an even deeper insight into the social hotspot popularity.

The system presents the overall sentiment of a social hotspot which can even be used for recommending points of interests to users as it can provide them with feedback for the popularity of a social hotspot in real-time. In Figure 3 a sentiment heatmap is depicted, where the red color represents negative, blue represent positive and green neutral Tweets.

Conclusion

In this paper, we have presented a system for detecting and identifying social hotspots from Twitter stream data, and applying sentiment analysis on the data. Utilizing the Twitter Streaming API, we have collected a significant number of Tweets from New York City and we have evaluated the quality of the retrieved data. Hierarchical and DBSCAN clustering algorithms have been analyzed for their usefulness in generating spatial clusters. We also elaborated on techniques for identifying social hotspots out of spatial clusters. Finally, we present an approach for sentiment analysis based on deep learning that is used to determine the attitude of users that participate in the hotspots.

Acknowledgments

We would like to acknowledge the support of the European Commission through the project MAESTRA – Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

Abdelhaq, H., Sengstock, C., & Gertz. 2013. Eventweet: Online localized event detection from twitter. In: *Proceedings of the VLDB Endowment (VLDB Endowment) 6*, pp. 1326–1329.

Birant, D., & Kut, A. 2007. ST-DBSCAN: An algorithm for clustering spatial—temporal data. *Data & Knowledge Engineering*, *60*: 208–221.

Bosch, H., Thom, D., Heimerl, F., Puttmann, E., Koch, S., Kruger, R. Worner, M., & Ertl, T. 2013. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *Visualization and Computer Graphics, IEEE Transactions on* (IEEE), *19*: 2022–2031.

dos Santos, C., & Gatti, M. 2014. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* Dublin City University and Association for Computational Linguistics, pp. 69–78.

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. 2013. Carmen: A twitter geolocation system with applications to public health. *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI),* pp. 20–24.

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*: 211–221.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, pp. 1746–1751.

Kisilevich, S., Mansmann, F., & Keim, D. 2010. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application.* 38.

Li, R-Y. 2013. A Study of Volunteered Geographic Information and Social Media (University of Calgary).

Liu, J. 2014. *A Location-Aware Social Media Monitoring System.*

Pak, A., & Paroubek, P. 2010. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining, LREC, 10*: 1320–1326.

Sui, D., & Goodchild, M. 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science, 25*: 1737–1748.

Walther, M., & Kaisser, M. 2013. Geo-spatial event detection in the twitter stream. *Advances in Information Retrieval*: 356–367.