Location Based Sentiment Mapping of Topics Detected in Social Media

Nasir Naveed¹, Muhammad Abbas¹, Zahid Rauf²

¹Department of Computer Science, Virtual University of Pakistan, ²Department of Electrical Engineering, Faculty of Information and Communication Technology, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta, Pakistan

Abstract
Social media connects people creating a virtual environment where there is no high up to dictate their posts. They feel free to appreciate or criticize the actions of others. Thus social media posts present true reflection of the public feelings over any social event. This sort of intrinsic social media content has great attraction for many stakeholders to explore public reaction on any incident or activity. This paper proposes a technique to visualize location based public sentiment on any event within specific locality. This research work harvests public posts from social media Twitter’s stream; extracts spatial, topical and sentimental information from the posts; performs statistical analysis and maps the results on geographic map. It provides stakeholders an unsupervised, quick and easy way to assess public opinion within a specific area.

Keywords: Social Media Analysis; Sentiment Mapping; Sentiment Classification; Data Mining; Social Media Topic Analysis

Corresponding author’s email: nasir@vu.edu.pk

INTRODUCTION
Social media has entirely changed communication ways. Once an event happens, instantly it reaches everyone on social media network. All this happens in real time; no time lapses are involved. It is because social media users, unlike conventional media, don’t require any sanction from their high ups; they just see and instantly share what they feel or experience without any hesitancy. That is why their posts present an absolute reflection of their minds on any topic. This is the only element that has turned social media into an attractive source of information for many stakeholders. Now people rely on social media reviews before making any daily life decisions. Some sharp stakeholders even use it for their reputation management, improving customer care services and maintaining market competition (Ribarsky et al., 2014). Literature review reveals that most of the past research work has been on sentiment classification (Asghar et al., 2014; Hao et al., 2011; Ahmed et al., 2014) or topic analysis (Naveed et al, 2011; Becker et al., 2011; Aellio et al., 2013; Sokolova et al. 2016) separately. Walther et al. (2013) worked on geo-spatial aspect of Twitter’s data while Lee et al. (2012) made contribution on spatio-temporal information of microblogs. This paper proposes a technique to combine spatial, topical and sentimental aspects of social media posts together and provide unsupervised, quick and easy way to assess public opinion on any topic within specific area bounds. This research work filters out Twitter’s stream for a specific area Pakistan and extracts 3 aspects from each tweet: i) spatial information on the basis of user profile location, ii) topical information through probabilistic model Latent Dirichlet Allocation (LDA) model and iii) sentiment information. The spatial information is used to map topical and sentimental information on geographic map.

MATERIALS AND METHODS
Proposed Architecture
The proposed architecture consists of four layers as shown in figure 1. First layer is responsible for dataset acquisition, while second layer performs dataset processing. Upper layers cover data access and visualization aspects of the processed dataset. The choice of Twitter as a source of our dataset is based on three reasons: i) Facebook users are mostly immature adults (school or college going) while in contrast Twitter users are serious and mature people, ii) Facebook follows friendship pattern where users need to be friends before they can follow each other while in contrast Twitter applies no such restriction. Its users need not to be familiar with each other. They can follow any one befitting their interest, iii) Above all, Twitter has soft policy for developers and provides access to its live stream publicly. Just, a simple authentication is required before accessing its live stream.
Twitter Connection
Twitter allows public access to its valuable contents with application programming interface API. It has been opted to use Twitter’s Streaming API for dataset collection within geographic bounds of Pakistan. The Streaming API offers real-time access to Twitter’s data. The only requisition is that the requesting application must be registered on Twitter which in turn requires a valid twitter account. It means only registered user can have registered application. For this purpose, an account has been created on Twitter. After the application is registered on Twitter’s page, it provides four credentials namely API key, API secret, access token and access token. These credentials are used by Open Authorization OAuth protocol to establish a Twitter connection.

Tweets Collection
After successful connection, a crawler application written in R language worked for about 4 months (from December, 2016 to April, 2017) and fetched 102,716 unique tweets. Twitter's stream was filtered out on the basis of three parameters: language, area and criteria. Only tweets in English language were considered because natural language processing NLP library lacked in functions to process other languages. In order to restrict tweets collection from within Pakistan, location attributes (i.e. latitude and longitude) of Streaming API were set in a way as if there was an abstract rectangle around Pakistan. Such a bounded rectangle was specified by south-west corner (62, 24) and north-east corner (74.5, 36) in terms of latitude and longitude. The third parameter track criteria was left blank. No @ or # tags were specified. This intentional decision of keeping track criteria blank was based on two motives; i) to realize the automatic topic detection feature of the proposed system in practice and, ii) to target all the tweets originating from within the bounding box across Pakistan.

The filtered tweets are by default saved to local JSON files. Since fresh tweets are by default appended to existing tweets thus JSON file size increases with the arrival of each new tweet. Definitely, JSON file size would have increased too much if all the tweets were pushed into single file. So, in order to keep JSON file smart, multiple sequenced JSON files were used for dataset collection. The R library rjson was used for this purpose. The JSON files were later parsed into data frame and transformed into .csv format for subsequent processing as illustrated by figure 2. The .csv files are relatively easy to manipulate. These .csv files became input to next higher level (dataset processing) of the proposed architecture.
Location Screening
Tweets’ location is a focal point of our research. Ideally, it should be the GPS location of source device but most of the users take the advantage of Twitter’s soft user policy and keep their GPS location off. Practically, it was observed that only 0.1% tweets carried GPS location. Obviously, such a frequently missing metadata cannot be a true representative of tweets’ location practically. So, alternately, user profile location seems to be a right choice for this research. It is available with almost all tweets but the only drawback is that it is manually set by user while making his/her profile on Twitter. It was observed that reluctant users managed to hide their location behind either partial information or even blank locations. A number of noisy tweets were also observed from adjacent countries like China, Afghanistan, Iran and India as well. The reason is clear; no bounding box can exactly fit over the irregular geographic shape of Pakistan.

Invalid tweets were screened out by using a 3rd party dataset WorldCities. It is a free as well as a reliable resource since many research scholars have used it in their research work. It contains information about 3173958 cities of 234 countries all over the world with 7 metadata data fields like Country, City, AccentCity, Region, Population, Latitude and Longitude. The screening process started with the removal of invalid location (empty string/blank) from the dataset. Next, partial matches were removed following a very simple rule that if a location contained some other country name it would be fully replaced by the country name. Thus, locations containing either the name of any neighbor country or its well distinguished city name were replaced by respective country name. As a consequence, dataset became free of all those tweets whose location was either missing or matching the name of some neighbor country’s name but, still, it contained some meaningless locations like invalid strings, celebrity names, funny phrases, etc.

For the removal of such invalid locations, we used reverse approach of keeping only those tweets which were related to only Pakistan; irrelevant or invalid locations were automatically knocked out. For this purpose, I developed an approach as illustrated by algorithm 1 (may be seen in appendix). Actual dataset WorldCities was reduced to just 349 pkCities by applying filter on its rows. Only those entries were selected where country name was Pakistan and population was greater than zero. Population condition was just to exclude small villages or localities from the dataset in order to keep things simple and boost up the speed of algorithm. Four columns including city name, region, latitude and longitude were kept. After that, each and every city from pkCities dataset was compared with the user profile location, if fully or partially matched, all the matched occurrences were replaced with it. Thus, all variants of a location like ‘lahore’, ‘lahore, punjab’, ‘lahore, pakistan’ and ‘lahore, the great’ agreed upon the same name ‘lahore’. Slang words like ‘lhore’, ‘faisal abad’, ‘lyallpur’, ‘minda bahauddind’, ‘rahim yar khan’, ‘mirpur’ ‘kiraanchi’, ‘k a r a c h i’, ‘gujarati’, ‘islamabd’ and ‘muzaffarbad’ were observed and corrected manually by using regular expressions. The entire process is shown in figure 3.
Text Cleansing

Tweets usually carry unstructured words that may mislead the sentiment as well as topic detection outcomes unless it is properly cleaned. For this purpose, basic techniques of natural language processing NLP were applied on the textual content of the tweets. These techniques include transformations like case lowering, the removal of RT, @, numbers, punctuation marks, URLs, stop words, white spaces, leading spaces and lagging spaces from text. Too short or too lengthy words were also eliminated by using regular expressions. Some raw tweets and their respective clean forms have been shown in table.

Table 1: Sample raw tweets and their cleanup version

<table>
<thead>
<tr>
<th>S.No</th>
<th>Raw Tweets</th>
<th>After Clean Up</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Good morning friends <a href="https://t.co/w6BB">https://t.co/w6BB</a> Nxpwyx</td>
<td>good morning friends</td>
<td>positive</td>
</tr>
<tr>
<td>3</td>
<td>@areeeshh I will understand trust me :) Try it!</td>
<td>understand trust try</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>@marvi_memon What will happen i can't wait.... :)</td>
<td>happen wait</td>
<td>Negative</td>
</tr>
<tr>
<td>5</td>
<td>humidity down 77% -&gt; 54%</td>
<td>Humidity</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

Topic Detection

After removal of the noisy content from tweets, next task was to identify what the text is about? For this purpose, we used CRAN package topicmodels that provided methods to run LDA model on dataset. LDA model is a probabilistic model that detects latent topics within textual part of the tweets. After a number of experiments, the model was fine-tuned by adjusting appropriate values of its various parameters like number of topics k, sparsity, burn-in period burnin, number of iterations iter, number of different starting points nstart and list of seed values each for a starting point seed. Initially, a document term matrix was constructed from a volatile corpus of clean tweets by using text mining package tm. Sparse terms were eliminated by adjusting sparsity level. Only 4317 terms could survive in the matrix for later processing at optimal sparsity level 0.9998. The 2nd important parameter k represents the number of most probable latent topics within a dataset. Unfortunately, LDA takes no responsibility about the value of k. It rather puts all the responsibility on user to decide how many topics could be the best ones for his/her dataset. It was really a tricky job to make an exact decision. Thanks to perplexity measure that helped us to reach an optimal value of k=100 for our dataset. Perplexity is actually the measure of uncertainty that how much a model gets surprised when it comes across some unseen data. Mathematically, it is geometric mean of per-word likelihood of held out dataset. Its minimum value gives the best value of k (Naveed et al., 2014). The perplexity variation with k
and the use of its optimal value during topic detection process is shown in figure 4 and figure 5 respectively.

![Figure 4: Perplexity variation with k topics](image)

We used Gibbs sampling method that performed random walk through the data. After a number of experiments, LDA model looked fine-tuned on burnin=1000, iter=200, thin=50 and nstart=3. LDA was run on document term matrix, which returned 100 topics. LDA model calculates probability score of each and every term for every topic and returns a set of most probable terms on the basis of decreasing probability score. The terms with highest probability appear on top of the list. We picked up top 5 terms of each topic like "university, killing, mashal, khan, mardan" as its representative terms. Later on, member tweets of each topic were identified on the basis of topic probability score distribution innately provided by LDA model. Thus, each tweet was associated to some or other topic with an exception of 4% tweets that could not be associated to any topic.

**Sentiment Detection**

Next step was to detect overall sentiment of each and every tweet. Sentiment means the overall semantic extract, emotion or feeling expressed by user through his/her words in tweet. We used RSentiment library for sentiment classification which originally returned 5 classes: very positive, positive, neutral, negative and very negative. These 5 classes were later normalized to only 3 classes: positive, neutral and negative by simply merging very positive and very negative classes into positive and negative ones respectively as depicted by figure 6. Thus sentiment class of each and every tweet was determined. RSentiment library follows Bag of Words Model which means that each word in the tweet has its own identity. The presence or absence of a word has no good or bad impact on sentiment value of other words. Each word has a sentiment score and the overall sentiment of a tweet is evaluated by taking the arithmetic sum of the sentiment scores of all the words in tweet.

![Figure 5: Showing the process of topic detection](image)
Visualization
Finally, the topical, sentimental and spatial information were integrated into single dataframe fully Processed Dataset and uploaded to google drive. This dataset was later accessed by a web-based application Shiny app to produce data visualization in form of graphs, word clouds and heatmaps as shown by figure 7. We developed Shiny app in RStudio environment and hosted it on Shinyapps.io, a free web hosting platform for Shiny apps. Our Shiny app consists of two components user interface UI and Server; UI makes the front end of the application while server works in the background to generate reactive response. UI has been divided into static and draggable panels. Static panel displays geographic heatmap on the basis of inputs selected from draggable panel and underlying dataset.

Figure 7: Showing visualization process
RESULTS AND DISCUSSION

Tweets Collection
Overall 102,716 tweets were collected during period Dec 29, 2016 to April 23, 2017. The tweets collection results as show in figure 8 reveal that 49.4% tweets came from Pakistan, 15.9% from nearby countries, 17.5% were invalid strings like <ed><U+ghs> and 17.1% were even blank. The reason of the tweets from nearby countries is the irregular geographic shape of Pakistan that does not fit into any bounding rectangle. The high cumulative fraction of blank and invalid locations indicates social media users’ reluctance to expose their identity. Figure 9 shows the region wise summary of tweets collection. Sindh, Punjab and Islamabad regions are on top of the list while KPK, Balochistan, Kashmir and FATA appear on bottom of the list. It shows high adoption of social media in top three regions while other regions are still far away from this race. It may be due to lack of internet services or low literacy factor there.

![Figure 8: Overall tweets distribution](image)

![Figure 9: Region wise tweets distribution](image)

Topics
LDA model returned 100 topics out of which 20% topics were clear. A few of them along with their short description are listed in table 2. The facts confirm that these topics remained in public interest during study period.

Sentiment
Figure 10 highlights two dimensions (topic, sentiment) of our dataset. It shows that maximum tweets were seen on Happy New Year event (11th topic). Overall positive sentiment remained dominant. Mashal Khan Murder (31st topic) and Syria War (81st topic) showed high negativity over social media. Other topics like Kashmir Freedom Movement (2nd topic), Electricity Load Shedding (3rd topic), Panamcase (26th topic), CPEC
Project (64th), and Disqualification of Nawaz Sharif (84th topic) claimed significant number of tweets. The results show that people in Pakistan show interest in national as well international issues. They support Kashmir Freedom Movement, Panamcase verdict and CPEC project but reject the events like unlawful killing of Mahsal Khan and the involvement of international anchors in Syria War. Overall fractions of positive, negative and neutral tweets were found to be 50.8%, 18.7% and 30.5%.

The summarized public opinion across top 20 profile locations is illustrated in Figure 11. It shows that during study period Twitter’s activity came mainly from three cities; Karachi, Lahore and Islamabad. The locations Pakistan and Punjab appeared because of partial information of profile location. Other cities like Rawalpindi, Peshawar, Hyderabad, Faisalabad, Multan and Gujranwala made a little contribution to dataset. The common reasons of the highest social media usage among top three locations may be good literacy rate and provision of internet services; population is another factor.

### Table 2: List of the most clear Topics Detected by LDA model

<table>
<thead>
<tr>
<th>Topic No.</th>
<th>Representative Terms of the Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>kashmir, india, indian, voice, freedom</td>
<td>Kashmir freedom</td>
</tr>
<tr>
<td>3</td>
<td>country, public, load, shedding, yet</td>
<td>Electricity load shedding</td>
</tr>
<tr>
<td>11</td>
<td>new, year, happy, wishing, eve</td>
<td>Happy New Year</td>
</tr>
<tr>
<td>26</td>
<td>court, panamaverdict, supreme, panamacase, decision</td>
<td>Panama case verdict</td>
</tr>
<tr>
<td>31</td>
<td>university, killing, mashal, khan, mardan</td>
<td>Mashal Khan murder</td>
</tr>
<tr>
<td>64</td>
<td>pakistan, power, cpec, china, end</td>
<td>CPEC project</td>
</tr>
<tr>
<td>77</td>
<td>team, cricket, play, misbah, captain</td>
<td>Misbah’s Retirement</td>
</tr>
<tr>
<td>81</td>
<td>syria, trump, attack, war, Russia</td>
<td>Syria war</td>
</tr>
<tr>
<td>83</td>
<td>khan, imran, ikleaderofnation, leader, vote</td>
<td>Imran Khan’s campaign</td>
</tr>
<tr>
<td>84</td>
<td>nawaz, sharif, minister, prime, gone</td>
<td>Disqualification of PM</td>
</tr>
<tr>
<td>96</td>
<td>pak, army, indian, must, death, Kalbushan (7th)</td>
<td>Kalbushan hanging</td>
</tr>
</tbody>
</table>

![Figure 10: Graph showing topic wise number of tweets for 3 classes of sentiment](image)

### CONCLUSION
The sentiment of topics detected in social media posts has been successfully mapped to their source locations on geographic map. One can easily visualize public sentiment on any topic coming from one’s own locality. Careful analysis of results reveals that social media has not been adopted by all regions of Pakistan equally.
Social media adoption in populated regions with good literacy rate is very high. Social media usage in less populated regions like KPK, Balochistan, Gilgit and FATA is very low. The possible reasons may be less population, poor literacy rate or lack of internet services there. Overall people think positively however they reject when there is a matter of unlawful activity. Their hearts beat with the Muslim community all over the world. They feel pain if any injury caused to Muslims anywhere in the world.

**REFERENCES**


Location Based Sentiment Mapping of Topics Detected in Social Media


