Potholes and Bad Road Conditions- Mining Twitter to Extract Information on Killer Roads

Swati Agarwal\textsuperscript{1}, Nitish Mittal\textsuperscript{2}, Ashish Sureka\textsuperscript{3}
\textsuperscript{1} BITS Pilani, Goa Campus, India
\textsuperscript{2} Noon.com, Dubai, UAE
\textsuperscript{3} Ashoka University, India

Research Track 5
The ACM India Joint International Conference on CoDS & COMAD
January 13, 2018
Goa, India
Motivation

• Increasing trend of adaptation of social media by Indian Government and public agencies to reach out to the people

• Public citizens use Twitter to post their complaints and report the incidents to the concerned authorities- *Citizensourcing*

• Complaints on killer roads contain the information about road irregularities and other issues causing high risks and discomfort to the citizens.

• To develop a system that can automatically identify complaint reports and overcome the challenge of manual inspection.

• To extract important information from tweets such as the exact geographical location which can be used to locate the fault
Examples - Bad Road Complaints

@nitin_gadkari pls re tar our village road, a farmer died y'day while jumping pothole, road condition is horrible, can't even walk, pls help

@Dev_Fadnavis @nitin_gadkari Ngar-Mnmd Hiwy. Frm LST 2 dys thr is big pothole at Nandgaon, nr Dehre Tollnaka, nrly 20-25 car's tyres damaged.

@nitin_gadkari Total chaos @NH4 near pashan PUNE No police to control busiest highway of India. Unplanned construction work does something SIR

@NHAllINDIA @nitin_gadkari big big pothole on service lane near hyatt gurgaon on NH8. Cars r getting stuck daily. Pls help in fixing it.

Travelling on NH58. Not a single street light between Ghaziabad and Merrut working on such a busy route @CMOfficeUP @nitin_gadkari

@nitin_gadkari on Hisar – Rohtak Highway huge Breaker is build near Sisar Cut & that too on the sharp turn jumped my car few ft save others
Related Work

• Mining online communications from Twitter to automatically determine road hazards by building language models. [Kumar et al 2014]

• Mining tweet texts to extract information on highways and arterial roads in Pittsburg and Philadelphia Metropolitan Areas. [Gu et al 2016]

• Analyzing real-time traffic related data for incident management. [Fu et al 2015] [Eleonora 2015] [Mai et al 2013] [Schulz et al 2013]

• Identifying traffic information on Twitter by mining traffic congestion, incidents and weather information. [Napong et al 2011]

• Examining the use of Twitter for making public awareness during heavy snow and public riots. [Panagiotopoulos et al. 2016]
Related Work - Steps Taken by Indian Government

- **TwitterSeva**
  - A program to address public complaints on one shared portal.
  - Uniting 200 social media handle
  - **Sub-units**
    - #DOTSeva, #BSNLSeva, #MTNLSeva, PostalSeva (Telecommunication and stakeholder departments)
    - #MociSeva (Ministry of Commerce and Industry)

- **Cybercell**
  - Union minister of Women and Child Development
  - Online complaints against women harassment
  - Online trolls and abuse against women
Novel Research Contributions

- First study on mining citizen's complaints and reports on killer roads posted on official Twitter handle of Government.

- We investigate the efficacy of spatial, contextual and linguistic features for identifying useful information from complaint posts.

- We build a text-analysis based model to enrich spatial features (geographical location metadata) in a tweet that can be used to discover insights from less informative reports.

- In order to conduct our experiments for this research, we create the very first database of citizens’ complaints on killer roads and highways and make our dataset publicly available for the research community so that our results can be used for benchmark, comparison and further extension.
Proposed Research Framework

Data Collection
- REST API
- Raw Tweets

Micropost Enrichment [1]
- #JointHashtags, Spelling Errors, Slang and Abbreviations, @username
- Enriched Tweets

Non-Complaint Tweets [1]
- Unknown
- Tweet Metadata, Named Entity Recognizer, N-gram modeling, Knowledgebase Graph
- Features Extraction

Problem Specific Identification
- Complaint Reports
- Complaint Tweets Classification
- One-Class/ Multi-Class Classification
- Features Vector Model

Empirical Analysis and Visualization

Experimental Dataset Collection

- Extracting all the tweets **posted to** the public agencies’ accounts
- Twitter Search API to extract the tweets consisting of **@username** in the tweets

- **Duration**: 8 weeks (18 July 2016 to 13 September 2016)
  - @MORTHIndia- Ministry of Road, Transports, and Highways
  - @nitin_gadkari- Union Minster of MORTHIndia

<table>
<thead>
<tr>
<th>Total #tweets</th>
<th>81,304</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
<td>17,511</td>
</tr>
<tr>
<td>Retweet</td>
<td>52,701</td>
</tr>
<tr>
<td>Replies</td>
<td>11,092</td>
</tr>
<tr>
<td><strong>English</strong></td>
<td>13,368</td>
</tr>
<tr>
<td>Unique</td>
<td>13,208</td>
</tr>
<tr>
<td><strong>Sampled</strong></td>
<td>3,302</td>
</tr>
<tr>
<td><strong>Users</strong></td>
<td>2,604</td>
</tr>
</tbody>
</table>
Features Extraction

- **Location** - to know the region (or city) from which the complaint is reported
- **Landmark** - the pinpoint location (or landmark) of the problem
- **Problem** - to know the type of issue faced by the citizens so that the complaint can be forwarded to the concerned department
1. Location

ED2- Tweets filtered by AISP classifier

“Mahatma Gandhi Marg”, “Vasant Vihar Colony”, “PVR Saket”, and “Nehru Place”
2. Problem Identification

- **Dysfunctional Facilities**- street light, traffic light, drainage

- **Risky and Hazardous**- pothole, street light, traffic light, hawkers, breaker, sign board, construction, intersection, diversion, animal, turn, crash, barrier

- **Poor Condition**- road, highway, traffic, flyover, underpass, bypass, chowk, expressway

- **Indiscipline and Carelessness**- barrier, repair, sign board, pothole, parking, police
2. Problem Identification

The diagram illustrates a process for problem identification, involving NLP and POS tagging. It starts with ED2 Core NLP, followed by POS Tagging. The next step involves converting a noun into an adjective form. The process then proceeds to a semantic similarity check against a lexical knowledgebase. Semantically similar terms are marked as related, while dissimilar ones are discarded.

The diagram includes:
- A data flow from ED2 Core NLP to POS Tagging.
- Conversion of a noun into an adjective form.
- Semantic similarity check against a lexical knowledgebase.
- High and low semantic similarity classifications.
- Related and discarded terms.
- Road-related distinct terms (C1, C2, C3, C4).

The terms are represented as follows:
- \(<W_1, W_2, W_3, W_4, \ldots, W_8>\)
- C1: W1 W2 W3...
- C2: W1 W4 W5...
- C3: W3 W5 W6...
- C4: W5 W7 W8...

This diagram outlines the methodology for identifying and categorizing related terms in a structured manner.
Classification

Unknown Tweets

- Irrelevant Tweets (IRT)
  - off-topic complaints
- Useful Tweets (UT)
  - complete reports
- Nearly-Useful Tweets (N-UT)
  - some information is missing

@nitin_gadkari @PMOIndia @narendramodi can we take some action on such wastage of public time n fuel as it's wastage of natural resources

@nitin_gadkari Sir ur attention pls. these r craters on Mum-Goa Highway near Chip lun. Being hopeful @Dev_Fadnavis @CMOMaharashtra

Street full of hawkrs no street light working whole city under chaos Ranchi has gone to worse please help @dasraghubar @nitin_gadkari
Enrichment of N-UT?

- Problem,
- Region,
- Landmark/Pin-point location

COMPLAINT TWEET POSTED BY THE USER

Rohit Sahai (रोहित)
@rohitsahai_

Major blunder by #DMRC at Huda City Center Metro Station. Entry to & Exit from to platforms clashing @nitin_gadkari

OPENSTREETMAP OUTPUT

Railway Station Shalimar HUDA City Centre Sector Road, Aardee City, Sector 52, Gurugram, Gurgaon District, Haryana, 122003, India

USER'S PROFILE

Rohit Sahai (रोहित)
@rohitsahai_

Joined February 2013

Gurgaon
Characterization of Complaint Reports

Distribution of Distinct Geographical Locations Identified in the Complaint Reports in our Experimental Dataset
# Experimental Results

<table>
<thead>
<tr>
<th>Tweet Type</th>
<th># Tweets</th>
<th>Overall accuracy: 67%</th>
<th>Predicted</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>N-UT</td>
<td>IRT</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>734</td>
<td></td>
<td>1088</td>
<td>131</td>
</tr>
<tr>
<td>Nearly-Useful (NC)</td>
<td>1668</td>
<td></td>
<td>376</td>
<td>569</td>
</tr>
<tr>
<td>Nearly-Useful (C)</td>
<td>50</td>
<td></td>
<td>254</td>
<td>34</td>
</tr>
<tr>
<td>Useful</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appreciation</td>
<td>166</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News</td>
<td>417</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information Sharing</td>
<td>97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,622</strong></td>
<td></td>
<td><strong>1718</strong></td>
<td><strong>734</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dysfunctional Facilities (C1)</td>
<td>402</td>
<td>345</td>
<td>282</td>
<td>44</td>
</tr>
<tr>
<td>Risky and Hazardous Roads (C2)</td>
<td>-</td>
<td>1142</td>
<td>933</td>
<td>145</td>
</tr>
<tr>
<td>Poor Condition Amenities (C3)</td>
<td>-</td>
<td>-</td>
<td>1042</td>
<td>92</td>
</tr>
<tr>
<td>Indiscipline and Carelessness (C4)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>252</td>
</tr>
</tbody>
</table>
Limitations

- API dependency
- Presence of multilingual scripts and text in a tweet
• We propose an approach for identification of complaints reported in road irregularities and bad road conditions.
• We propose to use the application of ConceptNet lexical Knowledgebase, NER, and geocoding location APIs to extract important components of a killer road complaint.
• Based on the available components in the tweets, we classify them into three categories: useful, nearly-useful, and irrelevant tweets.
• Our results shows that the proposed approach classifies these complaint reports with an accuracy of 67% and a recall of 65%.
• Maximum number of complaints are reported about the risky and accident prone roads while most of them are due to the poor condition of amenities.
• Complaints are reported from all over the India while maximum complaints are reported from Maharashtra, New Delhi, Uttar Pradesh, and Bihar states.
Future Directions

• Integration with other domains
  • Information Visualization
  • Metaphor analysis

• Modality and Dimensions
  • Multimedia content
  • Cross-platform analysis

• Novel Applications
  • Corruption barometer
  • Query vs. complaints
References

References


We are hiring full-time and visiting faculty in Computer Science

Our focus is on: AI, Networks, Systems, Software Science, and Theory (but get in touch anyway if you are doing exciting work in some other area of Computer Science)

Write to us: csis.office@goa.bits-pilani.ac.in
Tel: +91 832 2580 107
Thank you!
Questions?
BACK-UP SLIDES
Open Source Social Media Data

Twitter
- **Tweet**: Content and Blogger
  - #likes #retweets
  - URL, hashtag, media
- **User**: @username, profile name, bio information, location, and external URLs
  - Activity feeds, followers and followings
  - Tweets posted by the blogger

YouTube
- **Video**: Title, description, uploader, and related videos
  - #likes, #dislikes, and #comments received on the video
  - content and user of the comments and replies posted on the video
- **User**: Title and description of the channel
  - #subscribers and #subscriptions of a channel
  - featured channel, contact and subscriptions (optional)

Tumblr
- **Post**: Content, type, blogger and tags of a post
  - #notes, #likes, #dislikes and comments received on a post
  - who liked or reblogged a post
- **Blogger**: Title and description of the blogger
  - Activity feeds of the blogger
  - followers, following and likes of a blogger (optional)
OSSMInt Applications for Government

### Motivation

<table>
<thead>
<tr>
<th>Radical Users and Communities</th>
<th>Civil Protest and Unrest Forecast</th>
<th>Extremism and Hate Promotion</th>
<th>Religious Conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women and Child Online Abuse &amp; Harassment</td>
<td>Detecting Fake News on OSM</td>
<td>Public Citizens’ Complaints and Grievances</td>
<td>Online Recruitment in Radical Groups</td>
</tr>
</tbody>
</table>
Introduction

Citizensourcing

- disseminating information
- collecting complaints and grievances from citizens

Ministry of the Railway (@railminindia & @sureshpprabhu)

State Police (@delhinpolice & @mumbainpolice)

Ministry of Road and Transport (@morthindia & @nitin_gadkari)

State Traffic Police (@dtptraffic)

Income Tax Department (@incometaxindia)

Ministry of External Affairs (@sushmaswaraj)

- Department of children & youth affairs press office account (@DCYAPress)
- Department of health (@roinnslainte)

- Prime Minister Theresa May’s office (@Number10gov)
- National government (@GOVUK)
Case Study 1- Mining Public Complaints and Grievances

- General issues that are faced by public citizens and seek for immediate actions by the concerned authorities.

- Examples:
  - Corruption, illegal transactions from banks
  - No-refund on ticket cancellation, delay of train, poor facilities on train/station
  - Traffic violations
  - Crimes

- In 1st week of January 2017, 55 trains delayed in North India due to fog

- Heavy rainfall in Delhi-NCR caused massive traffic jam for hours
Case Study 1 - Mining Public Complaints and Grievances

@dtptraffic No one in Uttam Nagar follows traffic rule, no traffic police personnel is available. No enforcement, no fear. @AlokVermaCPDP

@IncomeTaxIndia My mom suppose to receive her #PANCard by 20June. Not received & No response from #FirstFlight. @bookcomplaint #FFCLisCRAP

@DelhiPolice 10000₹ wrongly deducted frm my ac in Lakshmi Nagar. can u help me in refunding the amount.

@RailMinIndia @sureshprabhu rain water is coming inside on seat through window sealing. Bedsheet n blanket both got wet. PNR NO. 8248914739
Case Study 2 - Mining Twitter to Extract information on Killer Roads

- **Bad Road Conditions**
  - Road Irregularities, Roughness, potholes, bumps, patchy surface and poorly designed speed breakers

- **Accidents Prone Roads**
  - Blind or improper curves, temporary diversions, traffic on under repair roads.

- **Dysfunctional Amenities**
  - Poor street lights or dim lights, encroachments on roads by shops, hawkers and broken road signboards

---

Motivation

- For motorists, death is lurking just around the corner
- According to Pune Traffic Police, one of the major reasons for accidents in the city is poor streetlighting during night.
Every Day, 400 People Are Killed In Road Accidents In India, Shows Government Data

400 road deaths per day in India; up 5% to 1.46 lakh in 2015

One serious road accident occurs every minute

16 die on Indian roads every hour

In 2015, Mumbai observed the most number of road accidents in the country
## Dataset Characterization

**Table 2:** Concrete Examples of Frequently Occurring 7 and 8 Character-gram Strings in the Experimental Dataset of Each Public Service Account

<table>
<thead>
<tr>
<th>Account</th>
<th>Hashtags</th>
</tr>
</thead>
<tbody>
<tr>
<td>@DelhiPolice</td>
<td>traffic, missing, abusing, arrested, detained, criminal, communal</td>
</tr>
<tr>
<td>@dtpTraffic</td>
<td>flyover, oddeven, parking, pillion, crossing, redlight, hospital</td>
</tr>
<tr>
<td>@IncomeTaxIndia</td>
<td>website, efiling, pending, invoice, property, interest, marriage, passport</td>
</tr>
<tr>
<td>@RailMinIndia</td>
<td>toilets, sleeper, medical, delayed, cleaning, security, drinking, stoppage</td>
</tr>
</tbody>
</table>
Case Study 2 - Mining Twitter to Extract Information on Killer Roads

**Motivation**

- **Road Irregularities**
  - @nitin_gadkari pls re tar our village road, a farmer died y'day while jumping pothole, road condition is horrible, can't even walk, pls help

- **Dysfunctional Amenities**
  - Travelling on NH58. Not a single street light between Ghaziabad and Merrut working on such a busy route @CMOfficeUP @nitin_gadkari

- **Under-construction Road**
  - @nitin_gadkari Total chaos@NH4 near pashan PUNE No police to control busiest highway of India. Unplanned construction work. do something SIR

- **Accident Prone Roads**
  - @nitin_gadkari on Hisar – Rohtak Highway huge Breaker is build near Siser Cut & that too on the sharp turn jumped my car few ft save others
Padding Space Correction

- Add a space after all special characters
  - Comma, semi colon, question mark, period, exclamatory mark

- Trim all extra whitespaces

@ nitin_gadkari Pls re tar our village road, a farmer died y’day while jumping pothole, road condition is horrible, cant even walk, pls help
Non-Complaint Tweets (AISP)

Shubham Singh
@iSinghShubham

Now this is what we call improvements in basic facilities. Kudos to @sureshpprabhu @RailMinIndia Surely #promiseinmotion

MORTHINDIA
@MORTHIndia

@nitin_gadkari invited to visit the State Transport Department Central Command Centre for Intelligent transport Systems #GadkarInUS 1/1

Amit Deoli
@amitdeoli

timesofindia.indiatimes.com/city/delhi/cyb ... main problem is under staff and stress. Police reforms needed for our @DelhiPolice. @CPDelhi
1. Frequent N-grams

- Grouped triplets of word n-grams frequently occurring in complains
- Addresses the challenge of keyword spotting method
- Use of WordNet lexical resource to map similar terms

<table>
<thead>
<tr>
<th>Account</th>
<th>N-Grams</th>
<th>Grouped Triplets</th>
</tr>
</thead>
<tbody>
<tr>
<td>@DelhiPolice</td>
<td>bribe, abuse, harassment, FIR, phone, action, report, complaint</td>
<td>{{bribe, abuse, harass}, {FIR, phone}, {action, report, complaint}}</td>
</tr>
<tr>
<td>@dtpTraffic</td>
<td>bribe, challan, violation, abuse, harassment, jam, commotion, congestion, accident</td>
<td>{{bribe, challan}, {violation, abuse, harassment}, {jam, commotion, congestion, accident}}</td>
</tr>
<tr>
<td>@RailMinIndia</td>
<td>train number, train name, coach, pnr number, bribe, corruption, report, complaint, action</td>
<td>{{train number, train name, coach, pnr number}, {bribe, corruption}, {report, complaint, action}}</td>
</tr>
<tr>
<td>@IncomeTaxIndia</td>
<td>pan number, ack, FIR, TIN, complaint, report, investigation, refund</td>
<td>{{pan number, ack, FIR, TIN}, {complaint, report, investigation}, {refund}}</td>
</tr>
</tbody>
</table>
2. Closed-Domain Keywords

- Terms occurring in a complaint specific to a department of public agency
- N-grams- to identify the type of the complaint
3. Events and Substances

- Concepts present in the tweets - not identified as the entities

- Extracted using IBM Watson Concept and Relationship Extraction API

- Uses various lexical resources and knowledge graphs as the backend databases (Yago, Dbpedia, WordNet, and Freebase)

Case Study 1

@dtptraffic @DelhiPolice no patrolling in #RaniBagh as loud music and crackers or on. No law following for crackers. #Noise @SatyendarJain

@sureshpprabhu @RailMinIndia : PNR number 8653150807... Passengers drinking alcohol n troubling other passengers
4. Location

- Location reported in the complaints
- Identify people, location, and organization
- Verify the location by applying geocoding on the entities
  - M. B. Road
  - Gandhi Nagar
  - AIIMS metro station
The bill given without #TIN (Tax identification number). When ask 4 TIN no he said he dnt hv TIN.@IncomeTaxIndia @sudhirchaudhary

@dtptraffic commercial vehicle plying during NO entry. Secondly driving in wrong direction. Creating huge traffic jam. Kindly tk action.
2. Problem Identification

- Big potholes on Delhi roads. It’s risky to drive at nights.
- Streetlights on highway not working
- Dim streetlights on highway.
- Highway is full of dark.
- There is complete blackout on highway.
1. Location - Landmark

- private
- administrative
- locality
- canal
- residential
- hospital
- hamlet
- aerodrome
- station
- trunk
- commercial
- construction
- neighborhood
- common
- bank
- motorway junction
- suburb
- kindergarten
- river
- bus stop
- bus station
- school
- restaurant
- unclassified
- motorway
- hotel
- secondary
- peak
- primary
- water
- public building
- bicycle parking
- service
- house
- stadium
- railway
1. Location- Region

- Village
- Town
- City
- District
- Territory
Enrichment of N-UT?

- Problem,
- Region,
- Landmark/Pin-point location

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Available</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mumbai contrast. Is this really the same city. Nitin Gadkari.</td>
<td>&lt;T_c&gt;</td>
<td>&lt;T_l, T_p&gt;</td>
</tr>
<tr>
<td>Nitin Gadkari even police vans do not follow traffic rules</td>
<td>&lt;T_p&gt;</td>
<td>&lt;T_l, T_c&gt;</td>
</tr>
<tr>
<td>Nitin Gadkari Sir NH 8 a black spot on your department? Seems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>you are feeling helpless because of Haryana Govt</td>
<td>&lt;T_l, T_c&gt;</td>
<td>&lt;T_p&gt;</td>
</tr>
</tbody>
</table>

**Figure:**
- **ED4:** Spatial Feature Enrichment
- **ED3:** Useful Tweets
- **ED1:** Sampled Data
- **Gaining Insights and Knowledge from Killer Road Complaints**
2. Problem Identification

\[ CS_{tK} = \]

<table>
<thead>
<tr>
<th></th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( \ldots )</th>
<th>( t_i )</th>
<th>( \ldots )</th>
<th>( t_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_1 )</td>
<td>0.4</td>
<td>0.5</td>
<td>( \ldots )</td>
<td>0.23</td>
<td>( \ldots )</td>
<td>0.50</td>
</tr>
<tr>
<td>( K_2 )</td>
<td>0.3</td>
<td>0.02</td>
<td>( \ldots )</td>
<td>0.7</td>
<td>( \ldots )</td>
<td>0.4</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( K_p )</td>
<td>0.24</td>
<td>0.25</td>
<td>( \ldots )</td>
<td>0.3</td>
<td>( \ldots )</td>
<td>0.7</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( K_m )</td>
<td>0.7</td>
<td>0.8</td>
<td>( \ldots )</td>
<td>0.6</td>
<td>( \ldots )</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Experimental Results

Case Study

Experimental Dataset

- Linear
- Polynomial
- RBF
- Cascaded Ensemble
- Parallel Ensemble

Precision

@dtpTraffic: 0.76, 0.75, 0.47, 0.56, 0.76
@RailMinIndia: 0.43, 0.42, 0.28, 0.33, 0.43
@IncomeTaxIndia: 0.62, 0.61, 0.36, 0.43, 0.83
Experimental Results

Table 5: Confusion matrix Results for C&G Tweets Classifiers (SVM with 3 different kernel parameters)- Tables illustrate the number of true positives, true negatives and false alarms generated by these classifiers.

<table>
<thead>
<tr>
<th>SVM Kernel</th>
<th>@dtpTraffic</th>
<th>@DelhiPolice</th>
<th>@IncomeTaxIndia</th>
<th>@RailMinIndia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>TN</td>
<td>FP</td>
<td>FN</td>
</tr>
<tr>
<td>Linear</td>
<td>184</td>
<td>294</td>
<td>58</td>
<td>178</td>
</tr>
<tr>
<td>Polynomial</td>
<td>170</td>
<td>296</td>
<td>56</td>
<td>192</td>
</tr>
<tr>
<td>RBF</td>
<td>182</td>
<td>146</td>
<td>206</td>
<td>180</td>
</tr>
<tr>
<td>Linear</td>
<td>188</td>
<td>471</td>
<td>252</td>
<td>123</td>
</tr>
<tr>
<td>Polynomial</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Experimental Results**

Table 6: Confusion matrix Results for Ensemble SVM Classifiers (Combining 3 SVM classifiers in Cascaded and Parallel Fashion)

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>@dtpTraffic</td>
<td>Serial</td>
<td>314</td>
<td>104</td>
<td>248</td>
<td>48</td>
<td>0.56</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Parallel</td>
<td>41</td>
<td>338</td>
<td>14</td>
<td>321</td>
<td>0.76</td>
<td>0.11</td>
</tr>
<tr>
<td>@DelhiPolice</td>
<td>Serial</td>
<td>41</td>
<td>63</td>
<td>545</td>
<td>2</td>
<td>0.07</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Parallel</td>
<td>6</td>
<td>536</td>
<td>72</td>
<td>37</td>
<td>0.08</td>
<td>0.14</td>
</tr>
<tr>
<td>@IncomeTaxIndia</td>
<td>Serial</td>
<td>46</td>
<td>11</td>
<td>60</td>
<td>6</td>
<td>0.43</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>Parallel</td>
<td>5</td>
<td>70</td>
<td>1</td>
<td>47</td>
<td>0.83</td>
<td>0.10</td>
</tr>
<tr>
<td>@RailMinIndia</td>
<td>Serial</td>
<td>253</td>
<td>206</td>
<td>517</td>
<td>58</td>
<td>0.33</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Parallel</td>
<td>41</td>
<td>659</td>
<td>64</td>
<td>270</td>
<td>0.39</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Dynamics of Religious Conflicts on Social Media
Term Obfuscation Detection in Secret Message Communication
Early forecast of civil protests

Data Collection
- Web Observatory
- MongoDB Query
- Local Terminal
- JSON Parsing
- MySQL Data

Semantic Enrichment
- TextRazor
- NERTagger
- SUTime
- Named Entity Recognition:
  Temporal, Spatial, Topic, People, Organization
- Event Related Keywords
- Learning Models:
  Crowd-Buzz/Commentary Planning/Mobilization

Event Forecasting
- C&C P&M
- Topic, Temporal Spatial NEs
- Performance Evaluation
- Chi-Square Distribution
- Frequency Model and Expressions Correlations
Online Radicalization Detection

- Hate promoting content identification
  - Twitter

- Radicalized users and communities detection
  - YouTube
  - Tumblr

- Intent-based radicalized and racist posts identification
  - Tumblr
Discussion on a variety of sensitive topics (religion and race), presence of hate promoting, radicalized and extremist content

Sensitive Topic Content

Expressing their Opinions
Shallow NLP Techniques
N-gram Feature Vectors Model
One-Class KNN and SVM
Linguistic and Contextual Metadata Analysis

Promoting Certain Ideology and Beliefs
Deep NLP and Semantics
Personality Traits, Writing Cues, Word Semantics, Emotions
Random Forest, Naïve Bayes, and Decision Tree

Users and Hidden Community

Users Uploading Extremist Posts and Forming Communities
Graph Traversal and Topical/Focused Crawler

Social Media Platforms and Data Sources

Discriminatory Features (Extracted and Derived)

OSSMInt Applications (Identification and Detection)

High-Level Approach

Requirement of Monitoring Social Media

Government Bodies Monitoring OSM
Future Directions

- **Novel Applications**
  - Corruption barometer
  - Intent vs. impact of hate promoting posts
  - Recruitment analysis in online radicalized communities
  - Fake news detection
  - NLP systems for security informatics
Micropost Enrichment Algorithm

- Consisting of data pre-processing and text enrichment techniques
- Addressing the problem of noisy data in tweets
Sentence Segmentation

- Find ill hashtags in the tweet (ignore the hashtag present in URL)

- Removal of filling terms
  - Umm, hmm, err,

- Trim consecutive special characters occurring together
  - ????, !!!!!, .....

@nitin_gadkari @SushmaSwaraj @narendramodi @vijayrupanibjp
#smartcitydahod#facing#worst#road#condition#
residence#ind
Joint Hashtag Expansion

1. **Common Separator (’-’ and ’-’)**
   - #no-strong-action-by-police = no strong action by police
   - #black_money = black money

2. **Uppercase Letters**
   - #MarchForDemocracy = March For Democracy
   - #CharchaOnRWH = Charcha On RWH
   - #FANTomorrow = FAN Tomorrow

3. **Alphanumeric String**
   - #TheriJoins100crClub = Theri Joins 100 cr Club
   - #NH8 = NH 8

4. **Longest subsequence**
   - #seriousissue = serious issue
   - #havesomesenseofchecking = have some sense of checking
   - #swachhbharat #swaranshatabdi #modisarkar
Spell Error Correction

- Given Sentence $S = \text{"plnty of ppl waiting frm past hour"},$

Acronym and Slang Treatment

• Domain Specific Slangs (fetched from frequent n-grams, n=1, 2, 3, 4)
  • NH (National Highway), Toll, RTO (Regional Transport Office), RC (Vehicle Registration Certificate), KM (kilometers), STN (station), Acc (accident) and RD (Road)

• Numeric
  • 2- to, two, too?
  • 4- for, four?

• Others (General slang or abbreviations)
  • B4- before, plz- please, idk- i don’t know, c- see
  • NoSlang online dictionary
Username Expansion

- Expanding @username mentioned directly in the tweet
- Extracting profile name using Twitter REST API

Ministry of Railways @RailMinIndia
Delhi Traffic Police @dtptraffic
Suresh Prabhu @sureshpprabhu
Proposed Research Framework

Data Collection
- REST API
- Raw Tweets

Micropost Enrichment
- #JointHashtags, Spelling Errors, Slang and Abbreviations, @username
- Enriched Tweets

Non-Complaint Tweets
- Unknown

Problem Specific Identification
- Complaint Reports

Complaint Tweets Classification
- One-Class/ Multi-Class Classification

Features Extraction
- Features
- Vector Model
- Tweet Metadata, Named Entity Recognizer, N-gram modeling, Knowledgebase Graph

Empirical Analysis and Visualization

Mining Public Complaints and Grievances on Twitter
Non-Complaint Tweets (AISP)

Now this is what we call improvements in basic facilities. kudos to @sureshprabhu @RailMinIndia Surely #promiseinmotion
Non-Complaint Tweets (AISP)

- Tweets consisting of an external URL
- URL not an image, video
- Quote tweets

Information Sharing

Amit Deoli
@amitdeoli

timesofindia.indiatimes.com/city/delhi/cyb ... main problem is under staff and stress. Police reforms needed for our @DelhiPolice. @CPDelhi
Non-Complaint Tweets (AISP)

- Tweets posted by other accounts or handle of same department or public agency
- Tweet posted by a verified account