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

Swati Agarwal
Birla Institute of Technology and Science (BITS), Pilani
Goa, India
swatia@goa.bits-pilani.ac.in

Nitish Mittal
noon.com
Dubai, UAE
nitishmittal94@gmail.com

Ashish Sureka
Ashoka University
Haryana, India
ashish.sureka@ashoka.edu.in

ABSTRACT
Research shows that Twitter is being used as a platform to not only share and disseminate the information but also collecting complaints from citizens. However, due to the presence of high volume and large stream data, real-time manual identification of those complaints is overwhelmingly impractical. In this paper, we identify the complaints and grievances posted on bad road conditions causing life risks, discomfort and poor road experience to the citizens. We formulate the problem of killer road complaints identification as a multi-class text classification problem. We address the challenge of keyword based flagging methods and identify several linguistic features that are unique for the killer road complaints tweets such as the issue reported in the complaint, pinpoint location of the issue, city or region location information. Our results reveal that not all complaint reports posted to Public agencies contain the sufficient information and are not useful. Therefore, we further propose a mechanism to enrich the nearly-useful tweets and convert them into useful reports. We present our results using information visualization and gain actionable insights from them. Our results show that the proposed features are discriminatory and able to classify killer roads complaints with an accuracy of 67% and a recall of 65%.

KEYWORDS
Bad Roads Complaints, Information Visualization, Lexical KnowledgeBase, Named Entities, Natural Language Processing, User Generated Data

1 INTRODUCTION
Bad road conditions such as road irregularities, roughness, potholes, bumps, patchy surface and poorly designed speed breakers make the road risky and hazardous for drivers resulting in accidents. Similarly, several crashes happen due to blind or improper curves, temporary diversions, traffic on under repair and under construction roads. Dysfunctional or dim street lights, encroachments on roads by shops, hawkers and broken or hard to see road signboards are also major causes of accidents. Bad road conditions based (the focus of the work presented in this paper) causes are different than accidents happening due to indiscipline or careless driving (driver’s fault such as speeding or drunken driving) or poor weather conditions (rain, storm, and snow). As per the statistics of Ministry of Road Transport and Highways (MORTH), Government of India, on an average 400 road deaths take place every day in India. According to MORTH statistics, in 2015, India had 501,423 road accidents causing 146,133 deaths while 10,727 people were killed in the crashes happened because of potholes. According to the reported accidents and newspaper statistics, Delhi recorded the highest number of road accident deaths (17 deaths per hour) in India in 2015. In the first quarter of 2016, over 300 deaths were reported in Uttarakhand state where accidents happened due to the sharp turns and absence of crash barriers. While, in Maharashtra, Meghalaya, Madhya Pradesh and Uttar Pradesh state, the maximum number of deaths happened due to the potholes and under-repaired roads and highways.

Previous studies shows that due to the 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. The complaints on killer roads contain the information about road irregularities and other issues causing high risks and discomfort to the citizens. Figure 1 shows concrete examples of bad road conditions complaints reported to the Indian Government’s official Twitter handle. Identification of such complaints reports on Twitter and resolving them in a timely manner is an important problem for the government and concerned authorities. However, the manual analysis of such complaints and gaining insights is overwhelmingly impractical due to the high volume and massive size of the data. Twitter statistics reveal that official twitter handle of government authorities and ministry of road, transport and highways in India receives an approximate of 6-10 complaint tweets per minute. Further, despite several efforts made by the government, in order to identify such complaints manually, many reports remain unaddressed. Kumar et al. [9] propose an approach to mine Twitter users’ online communication for identifying potential road safety hazards that pose driving risks. Gu et al. [6] propose a methodology to extract incident information on both highways and arterial roads of Pittsburgh and Philadelphia Metropolitan Areas. Fu et al. [5] describe an approach to extract and analyze real-time traffic related twitter data for incident management purpose. present a real-time monitoring system for detecting and analyzing the traffic incidents and congestion via Twitter communications. Schulz et al. [13] present a solution for allowing to increase the situational awareness by real-time identification of small-scale incidents. Panagiotopoulos et al. [12] present a study on examining the use of Twitter for making public awareness during emergencies such as heavy snow and public riots.

1http://www.thehindu.com/news/cities/bangalore
2http://indianexpress.com/article/cities/pune/
3http://morth.nic.in
4Accidents statistics in India
Motivation: The research presented in this paper is motivated by the existing literature and a need to develop a system that can automatically identify complaint reports and overcome the challenge of manual inspection. The key motivation of this paper is to extract important information from tweets such as the exact geographical location which can be used to locate the fault and present the information to the end-users of such applications (government agencies responsible for monitoring and repairing roads). However, the automated identification and analysis of such complaints is a challenging problem. Due to the presence of free-form text tweets do not have a defined structure or language format and hence are high likely to have grammar and spell errors. Further, the presence of multilingual texts and scripts in tweets, it is challenging to identify the linguistic features for building NLP based applications. Text classification or categorization and information extraction from tweets is thus a technically challenging problem. The specific research aim of the work presented in this paper is to investigate rule-based and machine-learning based text classifiers to identify bad road conditions reports. Our research aim is also to investigate information extraction and visualization to extract useful information and insights from the complaint reports.

2 RESEARCH CONTRIBUTIONS:

We conduct a literature survey in the area of mining Twitter data for identifying complaints and reports on bad road conditions. Based on our survey, we observe that most existing studies i) focus on mining users’ online communication on social media and news websites [9] [6], ii) building a prediction based model and situation awareness model for identifying real-time emergency situations such as heavy snowfall [12] and traffic issues [4] [14] [10]. However, in this paper, we mine the public citizens’ complaints reported on Twitter and extract the insights and information that are useful for the government. In contrast to the existing studies, the specific and key contributions of the study presented in this paper are the followings:

(1) To the best of our knowledge, the study presented in this paper is the first study on mining citizen’s complaints and reports on killer roads posted on official Twitter handle of Government.

(2) 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. We propose a content-disambiguation model to enrich our linguistic features for identifying exact problem reported in the tweets.

(3) We publish the first database of citizens’ complaints on killer roads and highways reported to official public agencies on Twitter. We make our dataset publicly available for the research community so that our results can be used for benchmarking, comparison and further extension.

We expect our results to be useful for the local government (Ministry of Road, Transport, and Highways) and concerned authorities (road, highway and traffic engineers) interested in identifying and fixing the causes of road accidents. Our analysis might help public agencies to perform data analytics on killer road reports and identify the geographical locations and issues in a time-efficient manner.
We use official Twitter REST API, we were able to collect a total of 81 weeks (from July 18, 2016 till September 13, 2016). Using Twitter we collect the tweets posted on these accounts in real time for 8

Table 1: Statistics of Experimental Dataset- Illustrate the number of Original and Unique tweets Collected After Data Extraction, Filtering and Sampling, Showing the Number of Sampled Tweets Consisting of Contextual Metadata (Image, Video, URL, Hashtag, @username mention). TS denotes the Timestamp (dd/mm) of the Tweet.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Total</th>
<th>Original</th>
<th>Re-tweet</th>
<th>Replied</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>81,304</td>
<td>17,511</td>
<td>52,701</td>
<td>11,092</td>
</tr>
<tr>
<td>Unique</td>
<td>13,368</td>
<td>13,208</td>
<td>3,302</td>
<td>2,604</td>
</tr>
<tr>
<td>Tweets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image</td>
<td>598</td>
<td>0</td>
<td>584</td>
<td>821</td>
</tr>
<tr>
<td>Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hashtag</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@mention</td>
<td>1673</td>
<td>8</td>
<td>20/07</td>
<td>Max 13/09</td>
</tr>
<tr>
<td>Lat-Long</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min TS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max TS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 EXPERIMENTAL SETUP

3.1 Dataset Collection and Characterization

In the work presented in this paper, we conduct our experiments on an open source Twitter dataset collected in real-time. We identify two official Twitter handle of concerned departments of Indian Government. @nitin_gadkari is the official account of Mr. Nitin Gadkari- current Union Minister of Road Transport & Highways and Shipping in India. @MORTHIndia is the official Twitter account of Ministry of Road Transport and Highways, Government of India. We use official Twitter REST API\(^1\) and search the tweets mentioning the screen name (along with @ symbol) of these accounts. Twitter REST API allows us to extract only past 7 days of data. Therefore, we collect the tweets posted on these accounts in real time for 8 weeks (from July 18, 2016 till September 13, 2016). Using Twitter API, we were able to collect a total of 81,304 tweets consisting of 17,511 original tweets, 11,092 replied tweets and 52,701 re-tweets.

In the work presented in this paper, we conduct our experiment only on English-language tweets. Therefore, we use "detect language"\(^8\) feature of Twitter API and identify the non-English-language tweets. Further, the aim of our study is to identify linguistic features of complaint reports and build a text classifier. Therefore, we also filter the tweets consisting of only images, videos, and URLs and no text-identified as undefined (und) using detect language feature of Twitter API. Table 1 shows the statistics of tweets remaining after filtering the non-English and unidentified language posts. In order to remove the bias from our data, we filter the duplicate records and perform a random sampling on our data to select a subset of tweets for our experiments. Table 1 reveals that after filtering and random sampling (25% of unique tweets) of records, there are 3,302 unique and English-language tweets posted by 2,604 unique users.

**Multi-media Content:** We observe that users embed external entities in their tweets to provide more details about the issue reported in their complaints. Therefore, in addition to extracting the textual content of tweets, we also download the available contextual metadata using Twitter REST API. We extract the available hashtags, @username mentions and count of URLs, images, and videos present in the tweet. We also extract the geo-location (latitude and longitude) of the users if enabled by the user while posting the tweet. Table 1 shows the number of tweets (after random sampling) consisting of various distinct contextual metadata. Table 1 reveals that only 18% (598 out of 3,302) of the tweets in our experimental dataset contain external images and 25% (821 out of 3,302) of the tweets contain distinct hashtags while only 8 out of 3,302 tweets contain geo-location information of the users posting these tweets.

3.2 Dataset Enhancement and Enrichment

In this Section, we preprocess the sampled dataset and address the challenge of noisy content in the tweets. We use the micropost enrichment algorithm proposed in our previous study Mittal et al. [11]. The proposed algorithm performs a syntactic enhancement of the tweets and consists of five phases primarily named as hashtag expansion, @username mention expansion, spell error correction, acronym and slang treatment and sentence segmentation. In this paper, we address the limitations of the previously proposed method and improve the accuracy of enrichment by varying the ordering of phases and number of iterations of some specific phases.

1. **Sentence Segmentation (SSN):** In the first phase of the micropost enrichment model, we enrich the syntactic structure of a tweet. We remove the filler terms (such as umm, hmmm, err) from the tweets. We also replace special characters appearing consecutively with one character. For example, ????? and !!!! are replaced with ? and ! respectively. Twitter does not identify the hashtags written consecutively or without spaces before "#". For example, a pattern of hashtags "#DwarakaExpressWay#DelhiRoads#Potholes", Twitter does not recognize any of the hashtags. We perform a cleaning on the extracted tweets and add a padding whitespace before every # and later trim all extra whitespaces occurring consecutively.

2. **Hashtag Expansion (HTE):** In hashtag expansion, we expand the joint hashtags used as a descriptive hashtag created by combining more than one words (alphanumeric strings or special characters). We use the method proposed in our previous study Mittal et al. [11] and split the long hashtag into disjoint words. For example, "#MumbaiPotholes", and #WakeupCM are split into "Mumbai Potholes", and "Wake up CM" respectively.

3. **Padding Space Correction:** Similar to the addition of space padding before hashtag (#), we add a whitespace before and after all special characters (comma, period, question mark, colon, semicolon, underscore and exclamatory marks). We further trim the consecutive extra whitespaces, fixing the presence and absence of spaces before and after the special characters. We, however, do not add whitespaces before the special characters appearing in @username mentions. For example, @poonam_mahajan.

4. **Spell Error Correction (SEC):** We pre-process each tweet in our experimental dataset and correct the possible spell errors. We use the application of Bing Search Engine AP for spell correction and extract the n-grams resulted as "Including results for". For a given sentence, 'pleas answr my query asap', we create a set of three 3-grams [n1: 'pleas answr my', n2: 'answr my query', n3: 'my query asap']. Here, 'answr' is replaced by 'answer' only if both n1 and n2

\(^1\)https://dev.twitter.com/rest/reference/get/search/tweets

\(^8\)https://dev.twitter.com/rest/reference/get/help/languages
Table 2: Concrete Examples of Tweets Recorded Before and After Executing the Micropost-Enrichment Algorithm Consisting of Five Phases (not in sequence)- User Mention Expansion (UME), Hashtag Expansion (HTE), Spell Error Correction (SEC), Acronym and Slang Treatment (AST) and Sentence Segmentation (SSN)

<table>
<thead>
<tr>
<th>SSN</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Umm nice. What about Vadra, Khurshid, Sheila Dixit... LoL. Where is the 370 pages report? @nitin_gadkari</td>
<td>@AamAadmiParty nice. What about Vadra, Khurshid, Sheila Dixit. LoL. Where is the 370 pages report? @nitin_gadkari @ArvindKejriwal</td>
<td></td>
</tr>
<tr>
<td>#InfrastructureMatters: @Nitin_Gadkari approves of #DelhiAmritsar e-way! #KnowMore <a href="https://goo.gl/Lk5uzn">https://goo.gl/Lk5uzn</a></td>
<td>Infrastructure Matters: @Nitin_Gadkari approves of Delhi Amritsar e-way! Know More <a href="https://goo.gl/Lk5uzn">https://goo.gl/Lk5uzn</a></td>
<td></td>
</tr>
<tr>
<td>It becomes very difficult 2 walk on #Delhi roads. Parents have given vehicles 2 their kids who knows nothing. 2de I saved myself 4 times @nitin_gadkari</td>
<td>@sureshprabhu learn from shipping where 1 accident why navigator leads 2 life time ban &amp; cancellation of license coordinate wid @nitin_gadkari</td>
<td></td>
</tr>
<tr>
<td>#InfrastructureMatters: @Nitin_Gadkari approves of #DelhiAmritsar e-way! #KnowMore <a href="https://goo.gl/Lk5uzn">https://goo.gl/Lk5uzn</a></td>
<td>MORTHINDIA Nitin Gadkari PMO India Narendra Modi pls take appropriate action. I ve raised so many complaints. No response whatsoever</td>
<td></td>
</tr>
</tbody>
</table>

are corrected as 'please answer my' and 'answer my query'.

5. Acronyms and Slang Treatment (AST): We expand the acronyms and slang present in our tweets in three stages: domain specific slang, standard slang, and user slang. In order to avoid the risk of noise in pre-processing, the algorithm proposed in Mittal et al. takes an input of domain specific acronyms. We do a manual identification on Twitter and identify several acronyms and slang commonly used in road and transport related tweets. For example, NH (National Highway), Toll, RTO (Regional Transport Office), RC (Vehicle Registration Certificate). If a term does not exist in domain specific slang keywords then we replace it with user slang. User slang are the slang that are commonly used on social media but are not present or recognized as standard slang used for those words. For example, PL is used for please whereas, it is defined as "parent looking" in a standard slang dictionary. We identify the 1, 2 and 3 n-grams present in the dataset and create an exhaustive list of such slang. In the third stage of slang and acronyms treatment, we replace the remaining slang with their standard definitions. We, however, do not replace any numeric character unless it is an alphanumeric string. For example, semantically, 4 can mean either word "for" or "four" while "r8" is used for "right" and "b4" is used for "before".

6. Username Expansion (UME): Expansion of direct mentions replaces the twitter @screenname with the user profile name making them easily recognizable by the named entity recognizers. For example, @Dev_Fadnavis, and @Ra_THORe are expanded to Devendra Fadnavis and Rajyavardhan Rathore.

In order to minimize the noise in output tweet, we apply domain specific slang conversion after every other phase of data enrichment. We observe that applying spell correction phase directly after hashtag expansion can rather create the noise in the data. For example, a hashtag #CMODelhi is expanded as "CMO Delhi" corrected as "COM Delhi" while CMO is an abbreviation used for "Chief Minister of". We use enriched tweets for feature extraction and classification and call them as Experimental Dataset 1 (ED1). Table 2 shows examples of killer road complaints tweets recorded before and after executing each phase of micropost enrichment algorithm.

4 FILTERING NON-COMPLAINT TWEETS

As discussed in Section 1, the official public service agencies account are open and anyone can mention them in their tweets. We observe that not all tweets posted on these accounts are complaint reports and rather are either off-topic or discussions not relevant to the complaint department. In order to classify a complaint report on bad road conditions, dysfunctional facilities or irregularities, we identify various features that are a strong indicator of a tweet to be a complaint report. However, we also observe that there are several discriminatory features that are a strong indicator of a tweet certainly not to be complaint report. Based on our manual inspection on official Twitter handle and non-complaint tweets identification model proposed in our previous study [11], we divide such tweets into 4 categories (AISP): Appreciation posts, Information Sharing and Promotional tweets.

Appreciation tweets are the posts made by citizens for praising the government for their work or resolving their previous complaints. Information sharing tweets are the tweets posted by users to share daily news about the events or policies initiated by the government. Further, due to the presence of several official accounts on Twitter, we categorize tweets posted by a different official account of same public agencies categorized as promotional tweets. Table 3 shows examples of appreciation, information sharing, and promotional tweets posted on official accounts of @nitin_gadkari and @MORTHIndia.
We use the AISP tweet classifier method proposed in our previous study [11] and classify AISP tweets and filter the unknown posts that may or may not be a complaint report about killer roads. We use these unknown tweets for further feature extraction and classification and call them as Experimental Dataset 2 (ED2).

5 FEATURES EXTRACTION AND SELECTION

We observe that all the complaints that are reported do not contain necessary and sufficient information about the issue faced by the people and hence are left unaddressed or unresolved despite considered by the authorities. Another fact that we come across is that unlike other department complaints, killer road related complaints can have fewer chances to get addressed based on the amount of information available in the tweets. For example, in a railway department, the following two complaints “no blankets are available in S3 coach of Hemkund Express.” and “Boarded Hemkund Express @PiyushGoyal @nitin_gadkari” are identified as persons’ names. Therefore, we use a combination of Named Entity Recognizer and GeoCoding API to extract the geographical entities mentioned in tweets. Figure 2 demonstrates the high-level framework for identifying a geographical location in tweets.

We use Indico Text Analysis API9 to extract the specific places and person referred in the tweets. Test Analysis model of Indico API allows a machine to discover a variety of knowledge from plain text input using Machine Learning and Deep Learning methods. Given a tweet \( t_i \in ED2 \) as plain text input, Indico API returns a list of JSON object of three key-value pairs. Each object represents a substring of locations and persons referred in the tweets. Test Analysis model of Indico API

9https://indico.io

Figure 2: Proposed Framework for Identifying the Geographical Location

Table 3: Concrete Examples of Non-Complaint Tweets- Classified into Four Categories: Appreciation (APP), Information Sharing (IS) and Promotional (PRL) Tweets.

<table>
<thead>
<tr>
<th>AISP</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>APP</td>
<td>Big big move @mansukhmandviya ji. Thank you for personally monitoring this. Thanks to @MORTHIndia and @nitin_gadkari ji also for this</td>
</tr>
<tr>
<td>IS</td>
<td>Commuters, get ready for more sun-kissed rides on the waters <a href="https://goo.gl/SleGW">https://goo.gl/SleGW</a> @PiyushGoyal @nitin_gadkari</td>
</tr>
<tr>
<td>PRL</td>
<td>Shri @nitin_gadkari lays foundation stones for 12 National Highways projects. #TransformingIndia</td>
</tr>
</tbody>
</table>

We discuss each of these features and the proposed method for their identification in the following subsections:

5.1 Geographical Location Identification

Twitter microposts are the user-generated data and do not have a defined structure for writing. Therefore, despite correcting the spell errors and performing sentence segmentation on tweets, we find that existing named entity recognizers are not able to identify locations in tweets with 100% accuracy. Further, we find that in Indian locations scenario, the majority of locations are identified as persons’ names. For example, for the locations “Mahatma Gandhi Marg”, “Vasant Vihar Colony”, “PVR Saket”, “Gandhi”, “Vasant” and “Saket” are identified as persons’ names. Therefore, we use a combination of Named Entity Recognizer and GeoCoding API to extract the geographical entities mentioned in tweets. Figure 2 demonstrates the high-level framework for identifying a geographical location in tweets.

We use Indico Text Analysis API to extract the specific places and person referred in the tweets. Test Analysis model of Indico API allows a machine to discover a variety of knowledge from plain text input using Machine Learning and Deep Learning methods. Given a tweet \( t_i \in ED2 \) as plain text input, Indico API returns a list of JSON object of three key-value pairs. Each object represents a substring of locations and persons referred in the tweets. In the input tweet, the confidence score varies between 0 to 1 representing the confidence of text analysis model for identifying it to be a place or person.

Using Indico API, for a given tweet input \( t_i \), we extract all the locations \( L = \{L_1, L_2, ..., L_n \mid 0 \leq i \leq length_t \} \) and persons \( P = \{P_1, P_2, ..., P_j \mid 0 \leq j \leq length_t \} \). Based on the score of verified locations, we divide the confidence score \( CS \) of locations into three categories- low \((CS \leq 0.2)\), medium \((0.2 \leq CS < 0.5)\) and high \((0.5 < CS)\). For each location \( L_i \in L \) with a confidence score \( CS_{L_i} \geq 0.5 \) and \( 0.2 \leq CS_{L_i} < 0.5 \), we filter them as \( L_{High} \) and \( L_{Medium} \) respectively. While we discard the entities identified as a place with a confidence score above 0.2. Table 4 shows examples of tweets and the confidence score of entities to be identified as location. As discussed above, due to the ambiguity in location names, we observe that many places are identified as persons and vice versa. Therefore, we discard the places having a medium confidence score but also identified as person names. The extracted geographical location entities for a tweet \( t_i \) are \( GL = L_{High} \cup (L_{Medium} - P) \). Further, if \( GL_m \in GL \) is a substring of \( GL_m \in GL \) then we discard \( GL_m \) and
filter only $GL_m$ entities identified as locations. For example, in the third example shown in Table 4, if "Vasant Vihar" is identified as a location then so are the "Vasant" and "Vihar". Therefore, we take the longest subsequence of the strings identified as locations and select "Vasant Vihar" as the location.

In order to identify the type of a location (region or landmark), we use OpenStreetMap (OSM) API\(^1\). OSM API takes geographical locations $GL$ as an input and identifies their geocode (latitude and longitude). Further, it also identifies the type of place (such as road, building, school, highway etc) based on the using the tags associated with their basic form. Each tag represents the geographic feature of the place. We group these features into two categories: region and landmark. We define a region as the name of a village, town, city, district and tertiary. Whereas, we define a place to be a landmark if it tells the pinpoint location, nearby area or locality. For example, geographic features such as school, church, hospital or building are labeled as landmarks. Table 5 shows the name of all amenities labeled as landmarks.

### 5.2 Problem or Issue Identification

As discussed above the public agencies accounts on Twitter are open and therefore, any one tag them in their complaints. However, during our inspection on Twitter, we observe that despite being a complaint report not all tweets posted on @MORTHIndia and @nitin_gadkari are related to killer roads. Further, in order to identify the complaint tweet, it is important to detect the information about topic or cause of the issue. For example, in a tweet "Big potholes on Delhi roads. It’s risky to drive at night!", identification of term "pothole" is important to know that the tweet belongs to a killer road category. Tweets are user generated data that contains free-form text and high likely to have different terminology for reporting similar complaints. For example, "streetlights on highway not working", "dim streetlights on highway", "highway is full of dark", "there is a complete blackout on highway" are reporting the same complaint using different terminology.

In our proposed method, we address the challenge of keyword based flagging methods for identifying the issues and problems reported in tweets. Figure 3 shows the high-level research framework for topic identification. We apply CORE NLP parser on the tweets identified as unknown after executing AISP classification (ED2). We perform POS Tagging on each tweet and identify the nouns, adjectives, adverbs and verbs. We extract the relevant terms and use them as an input to the Core NLP parser. We assign a confidence score to each term based on the parser output. We use a threshold to filter out the terms which do not meet the confidence score. We then use a lexicon to match the terms with the known terms. We also use a rule-based method to identify the entities. We use a rule-based method to identify the entities. We apply a rule-based method to identify the entities. We apply a rule-based method to identify the entities. We apply a rule-based method to identify the entities.

### 5.2.1 Table 4: Concrete Examples of Location and Person Names Identified using Indico Text Analysis Model- Illustrates the Entity and Confidence Score to be a Place or Person’s Name

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Place</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident at Wadkhal. Pls confirm if traveling towards Alibaug. Truck broke into 2 due to pothole.</td>
<td>&lt;Wadkhal:0.73, Alibaug:0.71&gt;</td>
<td>-</td>
</tr>
<tr>
<td>Pathetic highway since ages between Roorkee to haridwar. Corrupt ministry. I bet if you drive on own</td>
<td>&lt;haridwar:0.95, Roorkee:0.60, Pathetic highway:0.03&gt;</td>
<td><a href="">Roorkee:0.06</a></td>
</tr>
</tbody>
</table>

### 5.2.2 Table 5: List of All Map Features and Amenities Identified by OpenStreetMap API for the Tweets Present in Our Experimental Dataset and Labeled as Landmark

<table>
<thead>
<tr>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>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</td>
</tr>
</tbody>
</table>

\(^1\)https://wiki.openstreetmap.org/wiki/API_v0.6
Table 6: Key-terms Related to Various Issues Reported in Killer Road Complaints

<table>
<thead>
<tr>
<th>Related Terms</th>
<th>highway</th>
<th>flyover</th>
<th>animal</th>
<th>construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>[C1] street light, traffic light, drainage</td>
<td>0.1</td>
<td>0.34</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>[C2] pothole, street light, traffic light, hawkers, breaker, sign board, construction, intersection, diversion, animal, turn, crash, barrier</td>
<td>0.85</td>
<td>0.50</td>
<td>0.28</td>
<td>0.12</td>
</tr>
<tr>
<td>[C3] road, highway, traffic, flyover, underpass, bypass, chowk, expressway</td>
<td>0.73</td>
<td>0.4</td>
<td>0.31</td>
<td>0.03</td>
</tr>
<tr>
<td>[C4] barrier, repair, sign board, pothole, parking, police</td>
<td>0.26</td>
<td>0.53</td>
<td>0.35</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 7: A Snapshot of ConceptNet Distance Between Terms Present in Tweets and Issues Related to Killer Roads

<table>
<thead>
<tr>
<th>term</th>
<th>highway</th>
<th>flyover</th>
<th>animal</th>
<th>construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>street_light</td>
<td>0.1</td>
<td>0.34</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>pothole</td>
<td>0.85</td>
<td>0.50</td>
<td>0.28</td>
<td>0.12</td>
</tr>
<tr>
<td>intersection</td>
<td>0.73</td>
<td>0.4</td>
<td>0.31</td>
<td>0.03</td>
</tr>
<tr>
<td>hawker</td>
<td>0.26</td>
<td>0.53</td>
<td>0.35</td>
<td>0.12</td>
</tr>
</tbody>
</table>

adjective, and adverbs in our posts. We filter the terms that have already been tagged as places or persons during geographical location identification. In order to address the limitations of keyword based flagging methods, we compute the semantic similarity between the terms presented in a tweet and the terms related to road, transport, and highway (RTH) complaints. Research shows that a lexical database can be used to compute semantic similarity between two terms [3]. We use Conceptnet commonsense knowledgebase as a lexical resource and identify the conceptual similarity between the terms presented in the tweets and RTH complaints. ConceptNet is an ontology-based semantic network in which each node represents a concept and each edge represents the relation between two concepts. ConceptNet is an open source knowledgebase consisting of concepts extracted from common and informal sentences and daily basic knowledge [8]. It not only identifies the shortest path between two concepts but also identifies the common-sense associations that users make among the concepts. Therefore, due to the free-form structure of tweets, we select ConceptNet over WordNet and DBpedia lexical resources.

We conduct a manual inspection on Twitter handle of @MORTHIndia and @nitin_gadkari and identify several issues reported by public citizens about bad road conditions. Based on our observation, we define four categories of killer road complaints: [C1] Dysfunctional facilities, [C2] Risky and Hazardous (accident prone), [C3] Poor Conditions and [C4] Indiscipline and Carelessness. For each category, we define certain key-terms and compute their conceptual semantic similarity (using ConceptNet) with the terms present in the tweet.

Table 6 shows the list of all key-terms defined in the above-mentioned categories. Table 6 also reveals that there are several terms such as street light, signboard, and pothole that occur in more than one category. Therefore, we create a lexicon of all distinct words and compare them with the terms present in the tweet and identified as a Noun. The high confidence score of similarity computation shows that the tweet is high likely to be in the associated category. We use association method of ConceptNet to access the concepts and compute their semantic similarity. We use ConceptNet5.0 API and find the association between two given concepts using [assoc/c/en/concept1?concept2] request while we filter results computed for only English-language concepts using [filter=/c/en/]. Given a tweet \( t = \{t_1, t_2, t_3, ..., t_n\} \), a set of killer road complaints related key-terms \( K = \{C_1 \cup C_2 \cup ... \cup C_j\} \) where \( 1 \leq j \leq 4 \) and a network of knowledgebase \( N_{concept} \). We find a \( t_i \in t \) where \( POS_i \in \{NN, NNP, NNS, NNPS\} \), \( t_i \notin GL \) and \( t_i \notin P \) (refer to Section 5.1). Then, \( \forall K_p \in K \mid K = \{K_1, K_2, ..., K_m\} \), we compute the confidence score with each \( t_i \) as \( CS_{t_i, K_p} = \text{Conceptual Similarity}_{t_i, K_p, N_{concept}} \). Therefore, for each \( t_i \in t \) and \( K_p \in K \), we get a two-dimensional vector Score\(_{m,n}\) matrix.

If, \( \exists K_p \in K : CS_{t_i, K_p} \geq th \) then the topic of a complaint is said to be true for tweet \( t \). Whereas \( \forall K_p \in K : CS_{t_i, K_p} \leq th \), we convert the adjective form of noun words if available. If \( \forall t_i \in t \) and \( \forall K_p \in K \), the confidence score between concepts \( CS_{t_i, K_p} \leq th \), the topic of a complaint is said to be false for \( t \). Table 7 shows concrete examples of conceptual similarity score (confidence) score computed between various concepts- noun terms present in the tweets and topic-related keywords. For each \( t_i \), we select all \( K_p \) such that \( CS_{t_i, K_p} \geq th \) and assign various categories (Dysfunctional Facilities, Risk and Hazardous, Poor Condition and Indiscipline & Carelessness) to the tweet. \( \forall j, C_j = \text{TRUE} \) for \( t \) if \( K_p \in C_j \mid 1 \leq j \leq 4 \). We, however find several such examples, where the problem identification is challenging not only for the machine but also for human annotation. Such reports contains humor, ambiguity and sarcasm and increase the number of false alarms in problem identification. For example, “Nitin Gadkari Come to Gurugram! Enjoy the fun of Venice for free! Great offer by Haryana Government!”, “Nitin Gadkari NH17 is National Horror neglected by Govt. After Govt.”, and “But this highway is like a black spot on moon”.

6 CLASSIFICATION

As discussed above, Twitter users can direct mention a public agency account in their tweets while reporting a complaint. However, due to no restriction to the direct mention, users tag these accounts in many non-complaint and AISP tweets. Based on our inspection of available information in the tweets posted on @MORTHIndia and @nitin_gadkari, we define three classes of tweets: Useful tweets, Nearly-Useful tweets, and Irrelevant tweets. We use Rule-based classifier trained on the features extracted in Section 5 i.e. 1) problem

---

11http://conceptnet5.media.mit.edu
12https://wordnet.princeton.edu
Table 8: Concrete Examples of Reports Identified as Irrelevant Tweets (IRT), Useful Tweets (UT) and Nearly-Useful Tweets (N-UT) for Addressing and Resolving the Complaint

<table>
<thead>
<tr>
<th>Type</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT</td>
<td>can we take some action on such wastage of public time n fuel as it’s wastage of natural resources</td>
</tr>
<tr>
<td>UT</td>
<td>Sir your attention please. these are craters on Mum-Goa Highway near Chipilun. Being hopeful</td>
</tr>
<tr>
<td>N-UT</td>
<td>people are struggling with pothole roads in national capital Delhi, do something about that</td>
</tr>
</tbody>
</table>

or issue reported in the tweet, 2) city or town or region mentioned in the complaint and 3) landmark or pinpoint location of the complaint.

1) Irrelevant Tweets (IRT): We define a tweet as irrelevant if the post is not about the poor conditions and irregularities of roads or highways causing life risks, discomfort, hazard or poor experience to the citizens. Table 8 shows examples of such tweets posted on official accounts of Government authorities of Road and Transport department. Table 8 reveals that the tweet labeled as irrelevant is either off-topic or the authors discuss the problems related to road and transport; however the complaint is not about the poor conditions of roads or faulty and dysfunctional facilities.

2) Useful Tweets (UT): Useful tweets $U_t$ are the posts which are a clear indicator of complaints and can be used to identify the low-level details of the issue faced by the citizens. Given a tweet $t_i$ and a set of named entities $X = \{T_c, T_l, T_p\}$, we define $t_i$ as a useful tweet if $t_i \in N$ (dataset) and $t_i = \{X, T_o | \forall x \in X : P(x) \}$ where $P(x) \neq \emptyset$. While, $T_c$ denotes the name of city or region, $T_l$ denotes exact geographical location or landmark, $T_p$ denotes the problem or issue reported in the complaint and $T_o$ denotes other words in the tweet. Table 8 shows examples of tweet consisting of detailed information about the city, landmark location, and the problem reported in the tweet. For each $t_i \in U_t$, we refer them as Experimental Dataset 3 i.e. ED3.

3) Nearly-Useful Tweets (N-UT): Nearly-Useful tweets are the tweets posted for complaining a report but containing incomplete or insufficient information about the issue. For example, missing the exact location of the problem, ambiguity in reporting the issue, defining the problem on an abstract level and lacking the details. Given a tweet $t_i$, we define it as a nearly-useful tweet $t_i \in NU_t$ if $t_i \in N$ and $t_i = \{X, T_o | \exists x \in X : P(x) = \emptyset\}$. Table 8 shows examples of tweets posted as complaint reports but providing insufficient information. For simplicity of tweets, we removed the @username mentions from the tweets and used the corrected and enriched form of tweets (after applying micropost-enrichment algorithm). Table 8 reveals that in a N-UT complaint, the author has mentioned the problem of potholes in the city of New Delhi while the exact location of potholes is missing from the complaint. For each $t_i \in NU_t$, we refer them as Experimental Dataset 4 i.e. ED4.

6.1 Enrichment of Incomplete Reports

In this phase of proposed approach, we convert nearly useful tweets into useful tweets by extracting possible information about missing entities. As discussed above, a complete complaint report is comprised of three major components: city $T_c$ (or region), pinpoint location $T_l$, and the problem $T_p$ reported in the tweet $T$. However, based on the type of missing information, it is not always possible to convert an incomplete or nearly-useful tweet to a useful and complete report. For example, in the absence of issue expected to be reported in the complaint, it is hard to identify the problem component since the complaint is subjective and can be about any topic or it can be completely irrelevant as well. Therefore, we discard the tweets identified as nearly-useful tweets with no issue mentioned in the reports. We apply the geographical location hierarchy model (bottom-up) and use graph backtracking method to identify the region or locality for a given pinpoint location in the tweet. For a given pinpoint location $T_l$, we use OpenStreetMap API and extract the detailed information associated with a geographical location. For example, for HIT-Delhi, we are able to extract Govind-Puri Metro Station as landmark, Phase 3 as locality in Okhla region which is located in Delhi city. We use OpenStreetMap API to enrich the region and city information $T_c$ for each tweet consisting of only pinpoint locations $T_l$.

Since it is possible for two different locations to have the same name, we further use Twitter user location to verify the identified city information. We extract the location of all distinct users available in our experimental datasets 4 (ED4- Nearly Useful Tweets). If the city or state information mentioned on their Twitter profile matches with the location identified using OpenStreetMap API then we show the 100% accuracy of the location component enrichment otherwise we proceed with the enriched results without claiming the 100% accuracy in enrichment. For example, in a tweet "Major blunder by DMRC at Huda City Center Metro Station. Entry to Exit platform Clashing", "Huda City Center" is detected as a location entity. We use Using OpenStreetMap, we get the location or city name as Gurgaon which is also mentioned as the users’ location in his Twitter profile. Furthermore, for a given region or city, it is not possible to trace in the geographical location hierarchy model and identify the pinpoint location of the problem. For example, for the tweet "Mumbai can be termed as the pothole capital of India. Epic mess of @SwachhBharatGov.", the reported problem aims at the generalized picture of Mumbai roads and therefore it is infeasible to identify the exact location of the problem. While using OpenStreetMap, only state information can be extracted as Mumbai is a city and the capital of Maharashtra state which does not provide any information about the problem’s location.

7 EMPIRICAL ANALYSIS AND EVALUATION RESULTS

In this Section, we discuss the empirical analysis performed on the tweets, acquired experimental results and the characterization study performed on the complaint reports. We present the accuracy results of rule-based classifier and also discuss the influence of user-generated content on the overall accuracy of the proposed approach. We evaluate the accuracy of our classifier by comparing the observed results against actual labeled class. We conduct our
experiments on 3,302 random sampled tweets collected for @MOR-THIndia and @nitin_gadkari and report the accuracy of AISP and complaint tweets classifiers. Proposed AISP classifier identifies a total of 20.5% (680 out of 3302) tweets as certainly non-complaint (AISP) reports. Among 680 AISP tweets, 417 tweets are identified as news or information sharing tweets. While, 166 and 97 tweets are identified as appreciation and promotional tweets respectively. Based on our AISP experimental results, we perform rule-based classifier on the remaining 2,622 tweets and identify useful, nearly-useful and irrelevant complaint reports.

Table 9 reveals that a very small percentage of tweets (6.4%- 170 tweets out of 2622 reports) are identified as complete reports that contains all three important components of a killer road complaint. Whereas, the largest chunk of reported tweets is classified as incomplete or nearly-useful tweets (1718 reports out of 2622 tweets). Our experimental results reveal that further only a very small percentage of tweets are convertible (N-UT-C) into complete or useful tweets (50 tweets out of 1718 nearly-useful tweets) while 97% (1668 out of 1718) of nearly-useful tweets have either landmark or concrete problem component missing from the tweets and hence not-convertible. Table 11 shows the example of tweets present in our dataset and classified as nearly-useful tweets which further cannot be enriched or converted (N-UT-N). Table 9 reveals that due to the large percentage of reports with missing or incomplete information that is not possible to enrich, it is technically challenging to identify each and every complaint tweet efficiently.

In order to measure the performance of our classifier, we use standard metrics of Information Retrieval and compute the overall accuracy of the proposed approach. Table 9 shows the confusion matrix of the proposed classifier. Since the proposed approach is a multi-class classifier, we compute the performance of each class. Based on the results acquired by our rule-based classifier, we classify bad road related complaints with an overall accuracy of 67%. In addition to measuring the accuracy of our classifier, we also measure the overall recall value of the classification. Based on our results and as shown in Table 9, we achieve a recall of 65%. As discussed in Section 5.2, the complaints reported to public agencies’ accounts are user-generated content and lacks a standard format or terminology for complaining a report. Further, the excessive use of metaphor and sarcasm while reporting a complaint generates false alarms and impacts the overall accuracy of the classification.

7.1 Characterization of Complaint Reports

We perform a feature-based characterization on the tweets classified into 4 major categories of killer road complaints (Section 5.2):
from every state of India, there are some states from where the maximum numbers are reported. The map presented in Figure 4(b) shows that the cities of Mumbai, Delhi, Haryana, Bihar and Uttar Pradesh report complaints more relative to the cities of states like Karnataka, Telangana, Odisha or North East region of the country.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we present our study on identification of complaints reported in road irregularities and bad road conditions. We address the challenge of noisy and user-generated data by performing syntactic enrichment on raw-tweets. We also publish the first database of complaint reports on roads posted on official Twitter handle of Indian public agencies. We propose to use the application of ConceptNet lexical knowledgebase, named entity recognizers and geocoding location APIs to extract important components of a killer road complaint; such as problem reported in the complaint, landmark or pinpoint location, city or location information. Based on the available components in the tweets, we classify them into three categories: useful tweets (UT), nearly-useful tweets (N-UT), irrelevant tweets (IRT). We further propose a mechanism to enrich the N-UTs and convert them into UTs. Our results shows that the proposed approach classifies these complaint reports with an accuracy of 67% and a recall of 65%. We perform a characterization on complaint reports and our results reveal that 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. Further, the complaints are reported from all over the India while maximum complaints are reported from Maharashtra, New Delhi, Uttar Pradesh, and Bihar states.

Future work includes addressing the limitations of present study by improving the accuracy of location identification. Further, there are several complaints which contains humor, sarcasm and ambiguity. Such tweets does not provide information about the issue reported in the complaint. Our future work includes the use of metaphor analysis for extracting information from such ambiguous posts.

REFERENCES


ABOUT BEKASI IN PATH AND TWITTER. *HUMANIORA* 7, 3 (2016).


