To Call, or To Tweet? Understanding 3-1-1 Citizen Complaint Behaviors

Vanessa Frias-Martinez
College of Information Studies
University of Maryland, USA
vfrias@umd.edu

Abson Sae-Tang
Ecole Polytechnique Federale de
Lausanne, Switzerland
abson.sae-tang@epfl.ch

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
efm@tid.es

ABSTRACT

3-1-1 phone services offer citizens a crowdsourced platform to call and report about issues in their communities. In recent years, local governments and agencies have started to offer new communication channels through social media including Twitter accounts. Although not its main purpose, citizens are using these channels as a way to communicate 3-1-1 service requests to their local authorities. Furthermore, Twitter is also being used by citizens to share issues about their communities with friends and colleagues, without specifically addressing it to any local authority. In this paper, we analyze the behavioral similarities and differences between the use of formal 3-1-1 phone services and informal channels – like Twitter – to report about issues that affect a community. Our end objective is to help public institutions understand the relevance of informal communication channels as data sources for service requests. For that purpose, we design and evaluate a set of supervised classifiers that automatically label tweets as complaints and determine its type. A weighted multiclass classifier was selected based on its performance with precision and recall values of 86% and 62%, respectively. By comparing labeled tweets against official 3-1-1 phone service request records, we provide a large-scale analysis of citizen complaint behaviors over the two crowdsourced channels.

I INTRODUCTION

In 1996, the city of Baltimore (Maryland) was the first to deploy a *3-1-1* special phone number for non-emergency service requests. Citizens could call to report situations or complaints regarding a wide range of issues from traffic to noise or heating problems in their buildings. This initial deployment was run by the local police department and had an impressive initial success [22]. Since then, similar services have sprung out through the US and in many other countries, sometimes under different numbers. The most important contribution of 3-1-1 phone services is that they constitute a single point of contact to the different agencies that handle each type of service complaint. Instead of having citizens memorize individual agencies, they can just call and report a complaint that will be addressed in a transparent manner by the corresponding agency. For example, 3-1-1 can contact the Department of Transportation to address issues regarding real-time traffic, ferries or biking; similarly, they can contact the Taxi and Limousine Commission to solve issues that relate to taxis and VIP transportation services. But above all, 3-1-1 phone services share an important philosophy: to give tools to citizens to report about issues in their communities. As such, 3-1-1 phone services constitute an early form of crowdsourcing with citizens reporting non-emergency situations that matter at a local level.

With new times come new services. Local governments and agencies have adapted to the times and now offer new communication channels through, for example, Twitter accounts. These channels, which were originally intended to disseminate information and reach citizens, are not scrutinized in real-time and 24/7 by local agencies. However, citizens are increasingly using these channels as crowdsourced tools to report 3-1-1 service requests to agencies and local governments. Furthermore, individuals are also using their own social media accounts to report issues in their communities and share them with their followers and friends, often times without explicitly addressing them to an agency's account. As a consequence, given the popularity of social media and the increasing availability of smartphones, it becomes critical for agencies and local governments to understand how social media channels like Twitter are being used to communicate 3-1-1 complaints either directly to the agencies or indirectly to followers and friends. In fact, such analysis could prove determinant to evaluate whether city halls should follow and give service to social media channels in real-time, either with humans (like the 3-1-1 phone) or automatically through data analytics. In the future, fully responding to social media service requests might be the only way for local institutions to manage the millions of potential calls they would get without these alternative communication channels put in place.

In this paper, we focus our analysis on understanding the behavioral similarities and differences between the use of 3-1-1 phone services and Twitter channels to report issues that affect a community. The former constitutes the formal (official) channel to report complaints or service requests while the latter, although periodically screened, represents an informal (unofficial) channel. Our research

focuses on evaluating whether the nature of the channels has an impact on the way citizens report their concerns. There is an important body of qualitative and quantitative work that evaluates the use and social impact of formal 3-1-1 phone services in our communities [9, 20]. Similarly, many researchers have characterized Twitter users and their activity: detecting topic trends, predicting user features or classifying types of tweets, among others [11, 17, 19]. However, to the best of our knowledge, this is the first study that combines both channels with the objective of understanding how formal (3-1-1 phone) and informal (Twitter) crowdsourced channels are being used to report service requests to local authorities.

To compare official 3-1-1 phone service requests with Twitter activity, we first need to understand whether a tweet corresponds to a service request. For that purpose, we also present the design and evaluation of a set of text-based supervised classifiers that automatically determine whether a tweet is a service request and its type *e.g.,* traffic or noise, among others. By comparing labeled Twitter activity with 3-1-1 phone service requests over time, we can analyze large-scale behavioral similarities and differences between the two crowdsourced channels.

We expect that this analysis will be useful for local governments and agencies interested in understanding the relevance of informal communication channels as data sources for service requests. Our results might help city halls evaluate whether defining a more formal approach to data analytics from social media is necessary. Additionally, we also provide a tool (the automatic tweet classifier) to help organizations identify tweets that might contain relevant information for their organizations. Given that 3-1-1 services vary across cities, we focus our evaluation in the City of New York although our methodology would be valid for any other city with a similar service.

II 3-1-1 SERVICE REQUESTS

The City of New York created a 3-1-1 service for non-emergency requests back in 2003. As of today, it constitutes the largest non-emergency city service system in the US allowing citizens to call to 3-1-1 and report a complaint. Service requests expand across issues covered by over 40 different agencies in the city of New York. In fact, New Yorkers can report a wide range of issues from a complaint regarding a taxi driver (which is filed to the Taxi and Limousine Commission, TLC) to a noise complaint (taken care of by the New York Police Department, NYPD). Since 2010, these phone records are publicly shared through the city's Open Data Initiative [13].

More recently, the 3-1-1 service as well as most of the city's agencies have started to be present in social media through Twitter accounts, among others. These accounts allow agencies to inform citizens about a broad range of issues. However, such accounts have also started to be used by citizens as an informal forum to post complaints and comments without filling in any form but rather simply writing 140 characters. We refer to these communication channel as informal, because (i) it was not formally set up as a channel to post complaints and (ii) because the agencies clearly state that their Twitter accounts are not formally monitorized 24/7, as the 3-1-1 phone service is. In this paper, we want to understand the similarities and differences between these two distinct services: the *formal* 3-1-1 phone versus the *informal* Twitter interactions. Our aim is to analyze what type of complaints citizens report through each channel and to evaluate the behavioral similarities and differences between the two. Next, we describe the sources of information that we use to carry out such analysis.

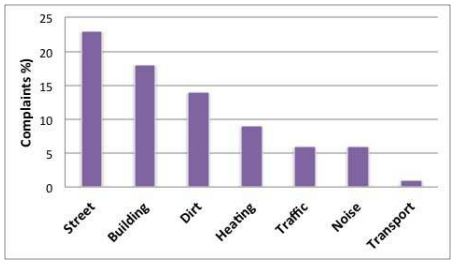


Figure 1: 3-1-1 Phone Most Frequent Complaints.

1 FORMAL 311 REQUESTS

The NYC Open Data initiative offers public access to all *formal* service requests (exclusively through the 3-1-1 phone) in the City of New York since 2010 [13]. A typical service request in the dataset contains, among others, the following information: (1) date and time of report, (2) department (agency) that handles the report, (3) complaint type, (4) descriptor with further details about the complaint and (5) street address with geolocation specified as a pair (latitude, longitude). In an attempt to understand types and volumes of complaints in the 3-1-1 service, we retrieved all service requests from 2013 and grouped them by type of complaint. Figure 1 shows the resulting statistics for the top most frequent complaints.

We observe that around 23% of the complaints refer to street or sidewalk conditions including situations like broken muni-meters, potholes or damaged trees covered by the Department of Transportation (DOT) and the Department of Parks and Recreation (DPR). The second largest volume of complaints (18%) contains service re-

quests regarding building issues like illegal work permits, structural problems or plumbing covered by the Department of Buildings (DOB) and Department of Housing Preservation (DHP). The third largest volume refers to dirt/sanitation which include complaints like rodents in public spaces, odors or dirty streets covered by the Department of Health and Mental Hygiene (DHMH). Other top complaints include issues regarding heating in buildings (although seasonal, the volume is really high); traffic complaints including illegal parking, which are managed by the New York City Police Department (NYCPD); noise complaints also covered by NYPD and Department of Environmental Protection (DEP) and requests regarding transportation services like the use of taxis or public transportation managed by the Taxi and Limousine Commission (TLC) and the Department of Transportation (DOT). Although these are not all the complaints gathered by the 3-1-1 service, they constitute over 95% of the total complaints thus representing a vast majority of the concerns reported by citizens in NYC.

For our analysis, we eliminate the heat building complaints due to seasonality *i.e.,* these complaints only appear during the winter season and only apply if the time period under study contains winter months. Additionally, we also eliminate building complaints since we want to focus our study on outdoor complaints regarding the city rather than indoor issues. Thus, the final set of complaint types that we use for our behavioral analysis are: street complaints (street); dirt complaints (dirt); traffic and illegal parking complaints (traffic); noise complaints (noise) and taxi and transportation complaints (transportation). Throughout the paper, we use this source of information to characterize *formal* complaints or service requests as opposed to *informal* ones collected from Twitter. See next section for details.

2 INFORMAL 311 REQUESTS ON TWITTER

In order to gather *potential* informal complaints or service requests from citizens in NYC, we consider two different types of tweets generated by the citizens themselves: (1) tweets addressed to any of the city agencies through their Twitter accounts (we will refer to these as @agency) and (2) tweets geolocalized in NYC (referred to as Geo throughout the paper). The first group of tweets (@agency) might contain concerns that citizens express, in an informal manner, to specific agencies in NYC like the Department of Transportation (@NYC_DOT), the 3-1-1 Department (@311NYC) or the New York Police Department (@NYPDNews). On the other hand, we collect the second group (Geo) to retrieve tweets from NYC that could be citizen complaints made in an informal manner to their followers and to the world at large. The purpose of this second group is to capture tweets from citizens who might be complaining without specifically addressing their complaints to any agency either because they don't know the Twitter ID or because they want to keep it more informal.

As opposed to the 3-1-1 phone records, both @agency and Geo tweets might or might not be service requests or complaints. In fact, while some tweets addressed to agencies are complaints, others will ask for information or praise a service. Similarly, not all geolocated tweets in NYC refer to service requests, on the contrary, very few will be city complaints. In the next sections, we describe the methodology and classification techniques used to determine which tweets are complaints (and its type) and to carry out the behavioral analysis comparing *formal* 3-1-1 phone complaints versus *informal* Twitter complaint behaviors. In terms of data collection, the @agency dataset is built using a set of track keywords containing the Twitter identifiers of each city agency as listed in [12]. We do not require these tweets to be geolocalized, although typically we observe that around 30% of them are. The Geo dataset is collected using a bounding box in the streaming API that covers all five NYC boroughs: Bronx, Manhattan, Staten Island, Brooklyn and Queens. It is important to clarify that if a tweet from @agency is geolocated and also appears in the Geo dataset, it is eliminated from the latter to avoid duplicates.

III METHODOLOGY

The focus of this paper is to analyze the behavioral similarities and differences regarding how citizens report various types of non-emergency complaints through formal or informal crowdsourced channels. As explained before, we focus our study on NYC and use as formal channel the 3-1-1 phone service where citizens can report service requests regarding the city. On the other hand, we consider two different informal channels: tweets addressed to specific NYC agencies (@agency) and tweets geolocated in NYC and addressed exclusively to followers (Geo). Specifically, we are interested in understanding the following three research questions:

- what are the behavioral differences across types of complaints between citizens that call to 3-1-1 and citizens that tweet to specific agencies?
- what are the differences across types of complaints between citizens that call to 3-1-1 and citizens that decide to tweet their complaints to their followers, without specifically addressing them to any agency?

- what are the behavioral differences between informal channels? do citizens tweet differently when they complain to agencies than when they share their complaints with their followers and with the world at large?

To analyze the differences between formal and informal complaint behaviors, we represent each of the three crowdsourced *complaint channels*: 3-1-1 phone, tweet to an agency (@agency) or geolocated tweet (Geo) as a vector where each component represents the relative volume (%) of a given complaint type with respect to all the other complaints for that complaint channel. By quantitatively comparing behavioral vectors over different time periods and qualitatively analyzing their content, we will be able to understand similarities and differences between channels. In order to compute each behavioral vector, we follow the methodology shown in Figure 2. First, to model the usage of formal crowdsourced complaints (3-1-1 phone), we collect all the complaints for a period of time T from the NYC Open Data Repository. Next, we use the complaint type field as a label to determine the type of complaint associated to each entry in the dataset. By identifying these labels in the open dataset, we can determine volumes of complaints and build the behavioral vector for the *formal* channel.

To compare these volumes against *informal* crowdsourced complaints made through Twitter, we collect tweets from @agency and Geo for the same period of time T as the *formal* complaints. However, unlike the 3-1-1 phone records, neither the @agency nor the Geo tweets are labeled by type of complaint. Thus, we propose to build a *complaint classifier* such that given a tweet from @agency or Geo, it determines whether it's a complaint and if so its type among the five under study. By gathering all the informal @agency and Geo tweets that were labeled as complaints, we can compute their behavioral vectors and compare these against the formal channel. Finally, Figure 2 also shows the methodology we follow to build the text-based complaint classifier. We propose a supervised approach that requires (1) a dataset with manually labeled samples to train the classifier and (2) an evaluation of different types of text-based supervised classifiers to select the one that best performs the classification task. By text-based, we mean that we use the content of the labeled tweets to associate vocabulary patterns to types of complaints. Once the classifier is built, it can be used with different labeled datasets to carry out multiple behavioral studies. The next two sections describe the labeling process, the classifier, its training and its evaluation.

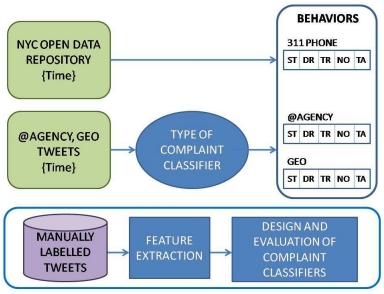


Figure 2: Methodology

IV LABELING TWITTER DATA

To build a text-based supervised classifier that determines whether a tweet is a service request (complaint) and its type, we first need a training dataset with a sufficiently large amount of labeled samples. However, the proportion of complaints about city services in Twitter is very low compared to the total volume of tweets a user generates. In fact, previous work using the text of tweets to detect flu outbreaks showed that the ratio of flu-related tweets to other types of tweets could be as low as 1:1000 [14]. Thus, randomly labelling tweets from a large dataset would probably result in identifying almost no complaints at all. To help us *smartly* select which tweets to label, we propose a two-stage process as shown in Figure 3. Step one builds a complaint/non-complaint Naive Bayes Classifier that determines the probability of being a complaint for any given tweet. Step two uses the output of such classifier to select as tweets to be labeled the ones whose probability of being a complaint is higher than a certain threshold. The underlying assumption is that labelling tweets that are determined to be highly probably complaints by the classifier is going to increase the chances of gathering a large amount of complaints as opposed to randomly labelling tweets which would probably result in the accumulation of large numbers of non-complaints.

To compute the complaint/non/complaint Naive Bayes Classifier (step 1), we select a two-week bootstrap sample (10thFeb-24thFeb) of Geo and @agency tweets (retweets –RTs– are not considered). Specifically, we select the tweets using a set of keywords that might be associated to service requests regarding street conditions, dirt, traffic, noise or transportation. These keywords, do not necessarily identify complaints but help us pre-select potential complaint candidates to be labeled. In fact, keywords like *rat* can refer to a dirt complaint *e.g., "I just saw a rat in the subway"* or to a study *"studies in rats show high levels of addiction"*. However, the posterior labelling process will disambiguate these two possibilities. We extract the set of keywords from the description field attached to the complaints in the 3-1-1 NYC Open Dataset (field

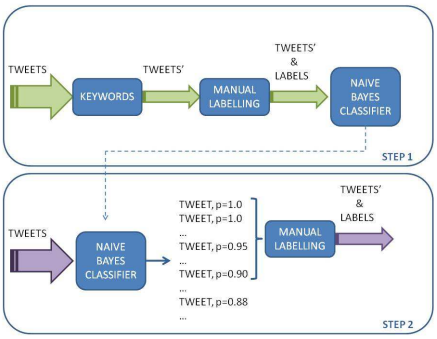


Figure 3: Naive Bayes Classifier.

four). Specifically, for each word, bi-gram and tri-gram in field four, we compute its frequency and χ_2 scores and select those with higher values in any of the two measures. Some examples of the keywords we have used for each complaint type are: for street, asphalt, pothole, road bump or muni; for dirt, waste, unsanitary, garbage; for traffic, blocked bridge, speed, highway; for noise, loud, party, restaurant and for transportation, overcharge, route, or taxi driver.

After applying the keyword-based selection to the two weeks of data, we obtain a total of 1799 tweets that we proceed to label manually. Our labelling process was carried out by three different labelers who participated in a training session to understand the complaints regarding city services and its types. Once labeled, we only used the tweets for which there was a 100% agreement in the final label. This resulted in a total of 320 labeled complaints and 1479 non-complaints. Next, we use these labeled tweets to train a Naive Bayes Classifier which, given a tweet, outputs its probability of being a complaint. The classifier, uses as features the uni-and bi-grams of the words that appear in each of the 1799 labeled tweets. For simplicity purposes, the words are all stemmed before the n-grams are computed. The final Naive Bayes classifier has an accuracy of 87%. Obviously, the quality of such classifier will not be optimal since it is built using a very limited number of labeled tweets that were pre-selected using a set of keywords. However, it will allow us to select – in an informed manner– larger amounts of potential complaint tweets to be labeled in step two.

Step two applies the complaint/non-complaint classifier to two months of Twitter data (Geo and @agency) collected from 1stMarch-1stMay. The complaint/non-complaint classifier outputs the probabilities of being a complaint for each tweet and we select the tweets whose probability is $p > 0.9$. This results in a dataset containing a total of 8814 tweets to be manually labeled as non-complaint or complaint and its corresponding type. To label all 8814 tweets, we worked with the same three labelers and only kept the tweets for which there was a total agreement in the final label. At the end of step 2, the total number of labeled complaint-tweets and its types was 1041 and the number of non-complaints 7773. Our final labels show that we identified an average of 150 complaints for each complaint type (street, dirt, noise and taxi/transportation) except for traffic complaints, which were the most frequent with a total of 426.

V COMPLAINT CLASSIFIER

In this section, we describe the design and evaluation of our Twitter complaint classifier using the text of the labeled samples obtained through the two-stage labelling process described in the previous section.

Corpus and Preprocessing Our corpus has a total of 8814 manually labeled tweets to train and test the classifier. We model each tweet as a bag-of-words (the position of the word is irrelevant) and maintain emoticons, hashtags and @'s as words. In terms of emoticons, we exclusively model the presence of either a positive or a negative emoticon in the tweet since these might express a sentiment associated to a complaint. Similarly, hashtags can be associated to relevant topics or words and the @ is kept to potentially model the agencies at which a tweet might be directed.

The underlying assumption is that if specific complaints are associated to a sentiment, a hashtag or an agency, these words will prevail across complaint tweets. As for the numbers that sometimes appear in the tweets, we pre-process them using regular expressions and only keep those which refer to an address, substituting the whole address with the word *<address>*. On the other hand, we eliminate the URLs and the check-ins from Foursquare that might appear in the tweets. We also eliminate common stopwords to make it easier to identify relevant complaint words. Next, we stem the words in the tweets using the English Porter2 stemming algorithm [6]. However, there are a few words that we do not stem to help us disambiguate between types of complaints and non-complaints. For example, we do not stem *park* and *parking* or *heat* and *heating*, among others.

After the pre-processing, we compute a document-term matrix (DTM) where each document is a tweet from the corpus and each term is either a uni- or a bi-gram of words that appear in the pre-processed tweets. Given the sparsity of the DTM *i.e.,* many terms appear only a handful of times, we associate to each term a weight and select as features exclusively those whose value is above a certain threshold. Specifically, we compute the weights as *tf-idf*

	Ensemble Classifier			Multiclass Classifier			Cascade Classifier		
Type	Precision	Recall	F1Score	Precision	Recall	F1Score	Precision	Recall	F1Score
Street	68.2 ±0.4	55.1 ±0.3	60.8 ±0.6	67.5 ±0.8	56.3 ±0.4	60.8 ±0.1	60.7 ±1.4	50.8 ±1.5	54.9 ±1.3
Dirt	73.2 ±0.8	59.3 ±0.8	65.2 ±0.1	74.8 ±0.1	58.7 ±0.3	65.2 ±0.4	68.2 ±1.1	52.7 ±2.9	66.4 ±1.7
Noise	75.4 ±0.3	57.8 ±0.1	64.3 ±0.8	75.1 ±0.1	57.9 ±0.4	64.3 ±0.6	69.0 ±0.8	54.2 ±0.8	66.6 ±0.6
Traffic	91.9 ±0.4	56.4±0.7	68.5±0.1	92.9 ±0.3	57.0 ±0.1	69.9 ±0.1	82.3 ±0.2	71.9 ±0.1	83.6 ±0.1
Transportation	82.6 ±0.2	61.3 ±0.8	70.1 ±0.5	83.4 ±0.5	62.7 ±0.8	71.7 ±0.1	73.7 ±1.7	60.7 ±1.2	68.4 ±1.3
Average	78.2	58.0	65.8	78.7	58.5	66.4	70.8	58.1	68.0
Baseline	64.2	46.3	55.5	63.7	45.5	56.1	59.2	51.1	54.8

Table 1: Average precision, recall and F1 Score for the Ensemble, Multiclass and Cascade classifiers.

–Term Frequency-Inverse Document Frequency – which measures the importance of each term proportionally to its presence in a tweet and inversely proportional to its presence in the whole corpus. We only select as features for the classifier the terms whose *tf-idf* is above 5% across the normalized weights in its class. This threshold allows us to reduce the sparsity of the features and to select those that will capture better the essence of each type of complaint. Thus, our final dataset contains 8814 labeled tweets characterized by a set of around $10,000$ features which are uni- or bi-grams whose tf-idf is above the 5% threshold.

Classifier Design We are interested in building a multiclass classifier that determines whether a tweet is a complaint and its type: street, dirt, noise, traffic or taxi/transportation. For that purpose, we design and evaluate three different types of classifiers: (1) a *Cascade Classifier*: consists of a two-stage classifier. Stage one is a binary classifier that determines whether a tweet is a complaint or not; followed by stage two which consists of a multiclass (5-class) classifier that determines the type of complaint for the tweets that were classified as complaints in stage one; (2) an *Ensemble Classifier*: a set of binary classifiers where each one is trained to differentiate between a non-complaint and one of the complaint types (street, dirt, noise, traffic and transportation). Given that we want to identify five types of complaints, we build an ensemble of five binary classifiers. The final decision is based on a weighted voting scheme: if only one classifier labels a tweet as a complaint, it is considered of that type; if more than one classifier label the tweet as a complaint, we assign the complaint type with a higher probability; alternatively, the tweet is labeled as non-complaint; (3) a *Multiclass Classifier*: a 6-class classifier that is trained to label as tweet as a non-complaint or as any of the five types of complaints. We implement each classifier with Support Vector Machines using an RBF kernel (SVM-RBF). The Cascade Classifier uses a binary SVM for the first stage and a multiclass SVM implemented with the one-versus-all approach (OVA), where each class is compared against all the others. The Ensemble Classifier uses five binary SVMs modified with Platt's methodology to compute the probability of each classifiers' decision [16]; and the Multiclass Classifier is also implemented with the OVA approach [18].

Classifier Evaluation In order to evaluate these classifiers, we train each of them using a 10-fold cross-validation over the dataset containing 8814 labeled tweets. Additionally, given that our sample dataset is imbalanced towards non-complaints (7773 vs. 1041), we undersample this class in each of the folds to equal the total number of complaints. In fact, undersampling approaches have been shown to improve precision and recall results in highly imbalanced sets [5]. Finally, 30% of the training folds are used to tune the best values for the C and γ parameters in the SVM. We discuss the performance evaluation of each classifier reporting its precision, recall and F1-scores since accuracy measures tend to be biased towards the majority class. Additionally, we also report the standard deviation for each measure across all runs in the 10-fold cv. Since we are interested in understanding how citizens complain in Twitter, we require both high precision and recall: a high recall to detect and analyze as many complaints as possible and a high precision to make sure that the complaints identified are real complaints. In order to guarantee that we extract *majority complaint behaviors*, we select the SVM parameters that yield recalls above 50%, whenever possible. For completion purposes, we compare the results against a *majority baseline* that associates to all samples in the dataset the label of the class that is more frequent: the non-complaint label in our analysis.

Table 1 shows the results for each type of classifier. The table shows that the best precision and recall values are achieved with both the Ensemble and Multiclass classifiers. In fact, both types of classifiers perform similarly with average precision and recall values of 78% and 58%, respectively. We also observe that both classifiers seem to perform the best for traffic complaints and the worst for street complaints. This result is probably associated to the fact that traffic complaints are easier to model than street

complaints which represent a broader range of issues from lighting to graffiti or fallen trees. Finally, when compared to the *majority baseline*, the Ensemble and Multiclass classifiers perform considerably better. On average, the precision of the baseline across the five classes is 56% and the recall 49%. Thus, our best classifier improves these results by 15% for the precision and by 12% for the recall. These differences were statistically significant at $p = 0.05$. On the other hand, the cascade classifier appears to perform a little bit worse with average precision and recalls of 70% and 58%, respectively (see Table 1). We hypothesize that these results are worse due to the high false positive rates associated to the binary classifier in the first stage. Exploring the associated ROC curve for the binary classifier in the first stage of the Cascade Classifier, we observe that true positive rate values above 70% are associated to false positive rates above 8%. This high rate, together with the imbalanced nature of the complaint/non-complaint distribution, funnels an important number of non-complaints to the second stage, which wrongly associates a type of complaint thus provoking a decrease in the precision values.

In an attempt to improve these results, while acknowledging the imbalanced distribution between complaints and non-complaints in our dataset, we test the performance of the best classifier (Multiclass) using SVM with class weights. This feature allows to associate a weight to each class in the dataset so as to penalize for misclassifications ı.e., the larger the weight, the larger the penalization if the sample is misclassified. For that purpose, we design a grid-based search to explore different combinations of class weights and select the one that outputs the best results in terms of precision and recall. Table 2 shows the best results obtained using a class weight vector in the Multiclass classifier. We observe that the average precision and recall increase considerably to 86% and 62%, respectively. The classes dirt and noise appear to be the ones that most improve from the use of weights (a 10% increase), which might be associated to the fact that these complaints are the ones with fewer number of samples in our dataset. In fact, SVM-weights improves the quality of the classifier for the classes that are more underrepresented in the training set. Although there is no direct work we can compare our results to, the most similar approach was presented by Liu *e*t al. [8]. The authors use lexical features extracted from user's statements on an online lending website to classify the categories of lender motivation. Since they report results using the $F0.5$ score, we need to convert our $F1$ scores into their corresponding $F0.5$ values. For that purpose, we use the precision and recall values in Table 2 with a $\beta = 0.5$. Our results show that our $F0.5$ score of 77.8 slightly improves their best value of 73.1.

Complaint Type	Precision	Recall	F1Score
Street	73.9 $\pm$0.2	57.0 $\pm$0.2	64.6 $\pm$0.1
Dirt	84.2 $\pm$0.1	66.2 $\pm$0.1	73.7 $\pm$0.1
Noise	88.1 $\pm$0.1	61.4 $\pm$0.3	72.0 $\pm$0.2
Traffic	97.2 $\pm$0.3	62.5 $\pm$0.1	75.9 $\pm$0.1
Transportation	87.0 $\pm$0.1	63.2 $\pm$0.1	72.5 $\pm$0.2
Average	86.1	62.1	71.7
Baseline	67.8	50.1	57.2

Table 2: Precision, recall and F1 Score for Multiclass Classifier with weights.

Finally, the dataset we have used to evaluate the classifier has a complaint to non-complaint ratio of ($\approx 1:7$). However, such ratio is probably not representative of what happens in the *real world*. For example, tweets regarding health events like the flu have been reported to be present in ratios 1:1000 [14]. Thus, in order to understand the impact that higher volumes of non-complaints –more approximate to reality– would have on the quality of the classifiers, we repeat the evaluation for two additional complaint:non-complaint ratios, namely 1:100 and 1:1000. Given that we have 1041 complaints, we need to select $100K$ and $1M$ non-complaints to build the new training sets with ratios of $1:100$ and $1:1000$, respectively. To do so, we use the Naive Bayes Classifier described previously. Recall that this classifier outputs the probability of being a complaint for any given tweet. Thus, by inputing two months of Twitter data into the classifier, we can select the $100K$ and $1M$ tweets with the lowest probabilities *i.e.,* higher probabilities of being a non-complaint. Our evaluation indicates that the precision, recall and F1 score values decrease less than a 1% when compared to the performance with the $1:7$ ratio. Thus, it is fair to say that, having larger volumes of non-complaints does not appear to decrease the quality of our classifiers.

VI CITIZENS' COMPLAINT ANALYSIS

In this Section, we analyze similarities and differences between formal and informal crowdsourced channels. For that purpose, we collect all the 3-1-1 phone complaints for the period of 25th May-10th June from the NYC Open Data Repository. We exclusively focus our analysis on the five types of complaints under study. These labels are already provided in the Open Data Repository. Additionally, we also retrieve tweets addressed to NYC agencies (@agency) and geolocated tweets in NYC (Geo) for the same period of time. It is important to clarify that these datasets are different from the ones we have used to train and test the classifiers, which were collected on earlier dates (10th February-1st May). Next, we use the

complaint classifier described in the previous section to label each retrieved tweet as a non-complaint or a complaint and its corresponding type. The resulting automatically labeled twitter complaint dataset and the 3-1-1 phone complaints are used to analyze behavioral differences across formal and informal complaint channels. As a sanity check, we also compared the 3-1-1 phone records with the 1799 tweets that we manually labeled (see step one in Figure 3). The results were similar to the analysis with the automatic labelling that we present in this section.

As stated earlier, we represent each of the three *complaint channels*: 3-1-1 phone records, tweets to agencies (@agency) and geolocated tweets (Geo) as a vector where each component represents the relative volume (%) of a given complaint type with respect to all the other complaints for that complaint channel. For example, the behavioral vector for the @agency tweets $[5, 10, 30, 20, 35]$ shows that citizens that tweet their complaints to an agency, mostly do so for noise and transportation complaints (30% and 35% of all the complaints respectively), whereas they tweet very little about street complaints (5%). The decision to work with relative volumes instead of the real number of complaints for each type and channel, is mostly due to the differences in complaint volumes across channels. In fact, we observe that the daily volume of 3-1-1 phone complaints in the period 25th May-10th June is ≈ 2800 while that of the Twitter channel is ≈ 400. Given that our complaint classifier has recalls of around 60%, in reality there might be a higher volume of informal complaints. However, it will still be much lower that the 3-1-1 phone volumes. This difference in volumes is probably due to the fact that the 3-1-1 phone is the official complaint channel as opposed to NYC agencies and personal Twitter accounts which are used in a more informal manner. Thus, since the comparison between channels' complaint volumes would be very much dominated by the formal one, we focus our research questions on understanding behavioral similarities and differences across complaint volumes relative to their own channel and compare these across channels. Additionally, to give more granularity to the analysis, we also construct separate behavioral vectors for weekdays and weekend days as well as for day and night time. Thus, for each complaint channel we define four different behavioral vectors: WD during weekdays; WE during weekend days; D during daytime (6AM-9PM) and N during nighttime (from 9PM to 6AM).

Figure 4 shows the percentages per type of complaint for each channel: 3-1-1 phone, Geo and @agency. The Jensen-Shannon divergence measure between the 3-1-1 phone and @agency is 0.19 compared to 0.14 between Geo and @agency, and 0.04 between 3-1-1 phone and Geo [7]. As a result, the largest behavioral differences appear between the formal 3-1-1 phone channel and the informal tweets addressed to specific agencies. On the other hand, the most similar behaviors are between the 3-1-1 phone and the informal tweets shared with contacts. Next, we analyze in depth the behavioral similarities and differences for each pair of complaint channels and finalize drawing a set of conclusions of interest for agencies and public institutions.

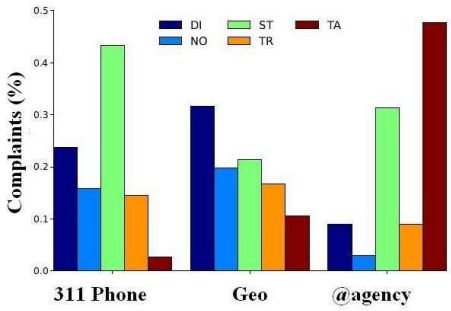
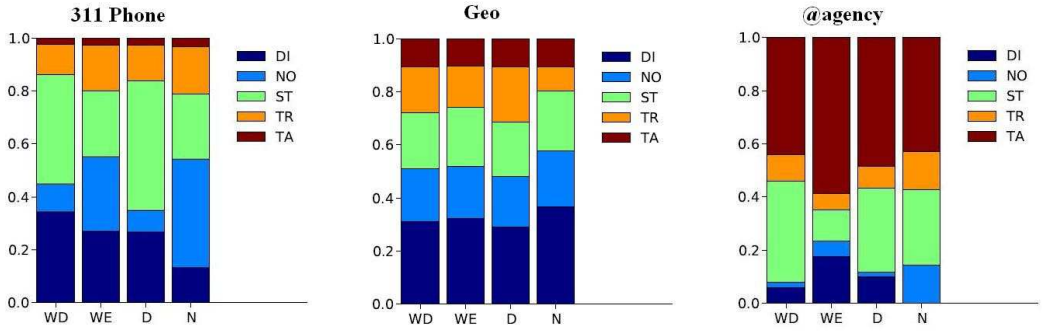


Figure 4: Distribution of complaints per channel (311 Phone, Geo and @agency) and type: dirt (DR), noise (NO), street (DT, traffic (TR) and transportation (TA).

311 Phone vs. @agency Focusing on the two most distinct crowdsourced channels, 3-1-1 phone and @agency, Figure 4 shows that while the 3-1-1 phone has high relative volumes for dirt and street complaints and average percentages across the others, @agency complaints mostly focus on street and transportation, with much smaller volumes for the other complaint types. A qualitative analysis of the @agency tweets labeled as transportation complaints shows that citizens tend to report complaints to the agencies' twitter accounts mostly when they have problems with taxis and taxi drivers. This means that, in relative volumes, citizens that use informal channels report more these transportation issues than citizens who use the 3-1-1 phone. We hypothesize that such finding is probably related to the scenario where the complaint takes place: while riding a cab. In fact, it appears that the informal channel provides a private (driver does not hear anything) and real-time communication between the citizen and the agency that the 3-1-1 phone does not offer. On the other hand, transportation complaints on the 3-1-1 phone are more varied including complaints regarding ferries, trains or airports which are almost inexistent in the @agency tweets.

Street complaints are reported in high relative volumes across both channels and typically describe issues regarding street lighting, muni-meters or damaged trees, complaints mostly addressed to the Department of Parks and



(a) (b) (c)

Figure 5: Distribution of complaints per channel (311 Phone, Geo and @agency), type: dirt (DR), noise (NO), street (DT, traffic (TR) and transportation (TA), and time of the day: weekday (WD), weekend (WE), day (D) and night (N).

Recreation (DPR). As can be observed, both formal and informal crowdsourced channels appear relevant for citizens to report this type of issues. On the other hand, while 3-1-1 also shows important relative volumes for dirt and noise complaints, these do not appear to be as relevant in the @agency tweets. We hypothesize that it could be due to one of the following facts: (i) the Twitter population is biased towards younger people who might cope better with or not as disturbed by issues regarding dirt or noise; or (ii) citizens are not familiar with the agencies' twitter IDs that deal with noise or dirt complaints, meaning that these agencies would probably need more publicity. In terms of temporal distribution, Figures 5(a) and (b) show the relative complaint volumes across days and times. We observe a common decrease in the relative volumes of street complaints from weekdays to weekends in both channels. However, we do observe an increase in the relative volumes of noise complaints also for both channels. A qualitative analysis of the tweets shows that these are mostly related to nightlife noise. Additionally, the @agency channel also shows an increase in transportation complaints during the weekends that does not appear in the 3-1-1 phone. As discussed earlier, citizens who might be traveling during the weekends probably run into taxi-related issues that they prefer to denounce in real-time, but with the privacy that the informal channel provides. Common to both channels, we also observe a logical increase in the volume of noise complaints from day to night time.

311 Phone vs. Geo As we discussed before, these two crowdsourced channels are the most similar with a Jensen-Shannon distance of 0.04. Figure 4 shows that, although in considerable different volumes, street and dirt complaints are the most popular in both channels. However, a qualitative exploration of the dirt complaints reveals an important difference between the two: while tweets labeled as dirt and addressed to followers mostly report disgust regarding the presence of rodents in the city (subway, restaurants, etc.), 3-1-1 calls represent a broader plethora of issues including trash collection problems or graffiti. We also observe that both channels are used for noise and traffic complaints. Although relative volumes are slightly higher for tweets than for the 3-1-1 phone, they mostly differ in the nature of their complaints. While citizens call to 3-1-1 exclusively to report city issues, Twitter users often times also seek to start a conversation with some of their friends and followers. For example, users might be compelled to complain to their friends about a party that is preventing them from sleeping at night *"that party downstairs is killing me, anyone up?"*; or to talk about the traffic jam they are stuck in *"stuck in traffic at GWB ahhhggg"*. In any case, although informally reported, these tweets contain information that could also be relevant to the corresponding city agencies.

There also exists an important difference in the relative volume of taxi/transportation complaints, which is almost triple for the informal channel. Again, it appears that the private and real-time nature of Twitter – tweets cannot be overheard by the driver or others – allows users to easily share their complaints while the event is happening. On the other hand, the 3-1-1 phone is probably used to complain after the event happened, which considerably lowers the probability of citizens actually reporting it. However, an interesting conclusion is that overall, it seems that citizens complain about similar things to the 3-1-1 phone and to their (Twitter) social networks. In terms of temporal distribution of complaints, Figures 5(b) and (c) show that the volume of 3-1-1 phone street and dirt complaints decreases from weekdays to weekends and from day to night, while traffic and noise complaints increase. However, we observe a slight opposite trend for the informal channel: citizens appear to tweet a little bit more about dirt complaints and a little bit less about traffic during

weekends and at night.

@agency vs. Geo Interestingly, the informal channels' behavioral patterns are quite different. Citizens appear to mostly tweet to agencies about transportation and street complaints whereas they share with their followers all types of complaints. It could be that transportation and street complaints are viewed as critical for the safety of the citizens, while other types of complaints, although important, are not as serious. For example, *"@nycparks Fallen tree at Elmhurst Av& 3rd should be removed or there'll be accidents!"* or *"@nyctaxi taxi driver with plate xxxx is driving and talking on the phone"* attempt to address more important issues than *"Ugghh Just saw a rat at 54th" or "I can't stand the party upstairs anymore!!!!!"*. On the other hand, citizens appear to be more prone to share dirt, noise or traffic complaints with their followers rather than with the corresponding agencies. As discussed earlier, many of the tweets addressed to followers and labeled as noise or traffic seem to seek a conversation with peers rather than to uniquely complain about a given situation. Thus, users appear to favor tweeting to their peers, rather than sharing the complaints with official agencies.

In terms of temporal differences between the two channels, we observe that the @agency behavior shares similar patterns with the 3-1-1 phone. In fact, we see an increase in the volume of noise complaints at night and during weekends together with a decrease in the street complaints for those same periods of time (see Figure 5(a)). On the other hand, as discussed before, the Geo tweets show slightly different trends, probably result of the nature of the complaints that also seek to start a conversation with their peers.

VII RELATED WORK

3-1-1 Service Analysis. Mazerolle *et al.* provide an in-depth assessment of the introduction of 3-1-1 services in the cities of Baltimore and Dallas [9]. In their analysis, the authors revealed a large reduction in the volume of 9-1-1 (emergency) calls that were transfered directly to 3-1-1 services mostly covering issues like traffic, parking or loud noise. Schellong *et al.* analyzed the use of 3-1-1 phone services during hurricane Wilma in the state of Florida [20]. Other studies regarding how citizens report service requests for their communities showed that citizens might choose different channels depending on what they want to report and their perception of safety or available time they might have. As a result, they suggest that 3-1-1 should have various types of input to offer a citizens an inclusive service [25]. Following the same philosophy, we believe that the analysis presented in our paper might prove useful for local governments interested in decreasing call volumes in 3-1-1 phone services by offering a real-time, 24/7 service on Twitter, similar to the one that is already being offered through the phone. Such approach would not only impact efficiency but would also increase government's transparency [2].

Twitter Content Analysis. Researchers have analyzed Twitter trends [11]; studied whether Twitter activity mirrors offline political sentiment [21, 23]; whether the content of the tweets can be used to infer geolocation of users [3]; how to differentiate between different types of users [15] or how to characterize and classify topics and their evolution over time in Twitter [17, 24, 26]. In our work, we focus on classifying tweets as complaints and its corresponding type. Since the volume of 3-1-1-type complaints is very limited, compared to other types of content, the work by Sadilek *et al.* and Paul *et al.* identifying flu-related tweets is also very relevant [14, 19].

Crowdsourcing Complaints. Other work, considers users as human sensors whose crowdsourced activity consists in monitoring a given service and to report failures or outages [4]. For example, Motoyama et al. [10] use Twitter to detect service problems in platforms such as Amazon or Gmail; and Agustine et al. [1] identify service complaints related to Netflix by analyzing tweets containing the word *Netflix* in almost real-time. However, these authors manually pick complaint samples from the feed to train the complaint classification systems. As opposed to these works, our approach does not preselect sentences but rather learns from large volumes of labeled data while handling unbalanced samples.

VIII CONCLUSIONS

We have presented a large-scale behavioral analysis to understand the similarities and differences between the use of formal (3-1-1 phone service) and informal (Twitter) crowdsourced channels to report service requests that affect a community. To carry out our study, we have also designed and evaluated a set of supervised classifiers that automatically determine whether a given tweet is a service request or not, and its type. Our analysis shows that a weighted multiclass classifier performed the best with precision and recall values of 86% and 62%, respectively. Our comparison between 3-1-1 phone service requests and labeled tweets show similar relative volumes of complaints between the phone service and citizens tweeting to their peers. However, tweets directed to the agencies mostly focus on transportation and street-related issues.

References

[1] E. Augustine, C. Cushing, A. Dekhtyar, K. McEntee, K. Paterson, and M. Tognetti. Outage detection via real-time social stream analysis: leveraging the power of online complaints. In *Proceedings of the 21st International Conference Companion on World Wide Web*. ACM, 2012.

[2] J. Bertot, P. Jaeger, S. Munson, and T. Glaisyer. Engaging the public in open government. *IEEE Computer*, 43(11):53–59, 2010.

[3] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Konowledge Management (CIKM)*, pages 759–768. ACM, 2010.

[4] V. Frias-Martinez, J. Sherrick, S. J. Stolfo, and A. D. Keromytis. A network access control mechanism based on behavior profiles. In *ACSAC*, pages 3–12, 2009.

[5] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[6] K. S. Jones and P. Willet. *Readings in Information Retrieval*. Morgan Kaufmann, ISBN 1-55860-454-4, 1997.

[7] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[8] Y. Liu, R. Chen, Y. Chen, Q. Mei, and S. Salib. I loan because...: understanding motivations for pro-social lending. *Proceedings of the fifth ACM international conference on Web search and data mining*, 2011.

[9] L. Mazerolle, D. Rogan, J. Frank, C. Famega, and J. Eck. Managing citizen calls to the police: An assessment of non-emergency call systems. In *National Criminal Justice*, 2001.

[10] M. Motoyama, B. Meeder, K. Levchenko, G. Voelker, and S. Savage. Measuring online service availability using twitter. In *Third Workshop on Online Social Networks (WOSN)*, 2010.

[11] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011.

[12] NYC. Official Twitter Accounts. `http://www.nyc.gov/html/misc/html/social_media`.

[13] NYC. Open Data. `https://nycopendata.socrata.com/`.

[14] M. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[15] M. Pennacchiotti and A. Popescu. A machine learning approach to twitter user classification. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[16] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers. MIT Press*, 2000.

[17] D. Ramage, S. Daumais, and D. Liebling. Characterizing microblods with topic models. In *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[18] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[19] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[20] A. Schellong and T. Langenberg. Managing citizen relationships in disasters: Hurricane wilma, 311 and miami-dade county. In *40th Annual Hawaii International Conference on System Sciences (HICCS)*, pages 96–96. IEEE, 2007.

[21] T. A. Small. What the hashtag? a content analysis of canadian politics on twitter. *Information, Communication & Society*, 14(6):872–895, 2011.

[22] Smarter. Governments: Center for technology in government. `http://www.ctg.albany.edu/publications/`, 2012.

[23] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, volume 10, pages 178–185, 2010.

[24] F. Villiers, M. Hoffmann, and S. Kroon. Unsupervised construction of topic-based twitter lists. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom)*, 2012.

[25] A. Woodcock, K. Frankova, and L. Garton. Voiceyourview: anytime, anyplace, anywhere user participation. In *Work: A Journal of Prevention, Assessment and Rehabilitation*, volume 41, pages 997–1003. IOS Press, 2012.

[26] H. Yulan, L. Chenghua, and A. Cano. Online sentiment and topic dynamics tracking over the streaming data. *Privacy, Security, Risk and Trust (PASSAT), International Conference on Social Computing (SocialCom)*, 2012.