https://doi.org/10.48009/3_iis_2019_175-185

Issues in Information Systems
Volume 20, Issue 3, pp. 175-185, 2019

POLICE USE OF SOCIAL MEDIA: COMPARING CLASSIFICATION METHODS

Kevin Mentzer, Bryant University, kmentzer@bryant.edu
Jane Fedorowicz, Bentley University, jfedorowicz@bentley.edu
Christine Williams, Bentley University, cwilliams@bentley.edu

ABSTRACT

With the advent of social media data, and in the pursuit of understanding meaning behind those data, text classification continues to grow in importance. Domain expertise is often needed to classify text effectively, but it is unlikely to be found in the tools that provide the means for unsupervised classification.

This paper examines a variety of text classification techniques applied to the domain of policing in the U.S. With police and their interaction with the public being in the public spotlight, understanding the messaging police send out via social media offers important insight into how the relationship between the police and the public evolves over time. However, classifying the tweets being sent out via the thousands of police forces across the U.S. is a daunting task. This work tests whether this classification can be automated, and explores the ramifications if it is.

Keywords: Policing, twitter, supervised classification, cluster analysis, social media

INTRODUCTION

The study of social media data often relies on understanding the meaning behind those data, so that methods for text classification continue to grow in importance. To classify text effectively, domain expertise is often needed, and often lacking from the tools that provide the means for unsupervised classification. In this paper, we demonstrate the efficacy of several text classification techniques by applying them in a single domain, police-generated social media messaging.

Most police departments in the United States use social media to share information with the public. However, it has been found that different police departments deliver different content through social media (Williams et al. 2018). It is still unclear what the impact this social media content has on relations with the general public: we have seen cases where a message intended to boost public relations in actuality backfired and heightened tensions between the police and the community (Tran 2014).

The first step in understanding the impact social media has on these relations is to understand how police departments actually use the medium. In (Williams et al. 2015; Williams et al. 2018), the authors manually classified the tweets of 5 different police departments to show how different police departments use Twitter for different means; for example, some police departments use Twitter to fight crime, while others do not. Expanding that work to include more police departments, or a longer period of time, becomes prohibitive with the manual classification technique.

This paper explores several categorization techniques to determine how effective an automated classification of tweets is when replacing manual classification. It provides insights into the best technique to employ, while also showing that higher levels of accuracy obtain for different types of tweets. This would allow future manual activity to focus on areas that are challenging to automatically classify, while letting the classification algorithm handle the easier cases.

Our key finding is that a simple rules-based technique outperforms more complex techniques. In particular, we find that just 107 terms, most of which are single words, are all that are needed to accurately categorize 76% of police tweets and some categories, including traffic and weather, with a 95%+ precision rate.

The paper is organized as follows: in the next section we explain why this domain is appropriate to study tweet classification. We cover the basics of text mining for those who are not as familiar with the process, and review the process of categorization of text. In section 3 we discuss the data used for our analysis and our preparation of the data for analysis. In section 4 we discuss the key measures that we use to evaluate each of the models. In section 5 we discuss our models and their results. In section 6 we discuss the findings more broadly and address what these findings mean to others interested in using this same technique. Finally, in section 7, we discuss the limitations of this work and suggest how it could be improved upon.

LITERATURE REVIEW

Police Use of Social Media

A 2016 Survey on Police use of Social Media by the International Association of Chiefs of Police and the Urban Institute, asked the question "What Does Your Agency Use Social Media For?" (International Association of Chiefs of Police 2016). Consider that for a moment. Social media use by police departments is public information, yet researchers are still asking how police departments are using Social Media. Studies such as (Edlins and Brainard 2016; Fernandez et al. 2017) show that while researchers commonly use the metadata associated with tweets including retweet count, author location, number of favorites, datestamp, etc., with the exception of perhaps measuring overall sentiment, few analyze the actual text of these social media postings, because of inherent domain-specific complexities.

With approximately 15,400 police agencies in the United States (U.S. Department of Justice 2015), nearly all of which are using social media (Edlins and Brainard 2016), obtaining some level of understanding about those tweets is extremely challenging. Even if each of those agencies uses social media for limited purposes, it is still an extremely large number of social media posts that would have to be classified to answer the question "What Does Your Agency Use Social Media For?".

Text Mining

We propose to perform text mining on the tweets themselves to better understand the topics of discussion being employed by various police agencies. Text mining begins by breaking down each sentence, and each word, to derive its part of speech, the role or entity the word plays, and understand the root of the word so that case or tense differences are eliminated, and grouping words are compiled into synonym lists. The terms are then filtered down to reduce dimensionality by eliminating terms that would add little or no value (such as the words "the", "as", and "do").

Different techniques have been proposed to attempt to classify the text. One technique is clustering, to aggregate the documents based on similarities, by taking the remaining words to create a term-by-document matrix. The general nature of this problem results in a matrix that is inherently quite sparse meaning that most terms only appear in a very limited number of tweets. The sparsity of the data make it extremely difficult to analyze computationally, and as such, some further reduction in dimensionality is desirable. Singular Value Decomposition (SVD) is often used to reduce this dimensionality by transforming the sparse matrix into a more compact form (Sudhan et al. 2012). While a low number of dimensions (i.e. < 50) are useful for general clustering, higher numbers (50-200) are necessary for classification of text (Sanders and DeVault 2004). The result of this technique means that documents that are similar will aggregate in clusters while minimizing similarity of documents across clusters (Sudhan et al. 2012).

A second technique that could be employed is a rule-based model. In a rule-based model the objective is to build a set of simple rules using the words available, to accurately predict a pre-defined category. Unlike clustering, the documents, or some subset of the documents, have to have been classified already. This classification becomes the target. The rules are built iteratively with the most successful single rule being set first. Subsequent rules, for that target, build upon the earlier rule. Caution has to be employed to not overfit the data.

Tweet Classification

Tweet classification is challenging. On the one hand, one could say the character limitation (140 characters at the time the tweets in this study were pulled, more recently raised to 280 characters) keeps the vocabulary small and, as such, may minimize the sparsity within the term-by-document matrix. On the other hand, this character count lends itself to creative spelling at times and the use of abbreviations (Kouloumpis et al. 2011), which could lead to increases in sparsity or a higher level of effort required to identify synonyms.

Tweet classification also must address non-text data. For example, consider the common phrase, "a picture is worth a thousand words." In text analysis, we are left with an image and no words. In addition, the rising use of emoticons to convey feeling also poses a challenge to those studying Twitter (Kouloumpis et al. 2011).

While sarcasm and irony have been of high interest to those who study tweets (González-Ibánez et al. 2011), the fact that these are professional police accounts leads us to believe that these language expressions appear too infrequently to warrant special consideration.

Prior studies have shown that a high accuracy level for categorization of documents could be obtained with a few or even a singular word (McCallum and Nigam 1998) when the text is not highly complex. Because of the character limitation, tweets tend to be simplistic in nature. Others (Zubiaga et al. 2015) have shown that hashtags can be a rich source of data to identify Twitter trends.

The wide-ranging options affecting the choice of text classification methods leads to our research question:

RQ: Which classification techniques most accurately categorize police tweets?

DATA AND METHODS

The data for this study were obtained from the authors of (Williams et al. 2018) and include 2,090 tweets from 5 police departments in Massachusetts, USA. The tweets were pulled from the 5 primary police Twitter accounts over a 3-month period (May 1st through July 31, 2014). The tweets were manually classified using an open coding approach which resulted in 10 different categories of tweets. This manual classification was necessary to provide a baseline against which to measure our results. Figure 1 shows the breakdown of tweets by category. Our data differ slightly from (Williams et al. 2015) in that there were some tweets that the authors did not use that we classified as unknown, and we combined the Traffic and Accident categories into a single Traffic category because of close overlap (this will be discussed further in the results section).

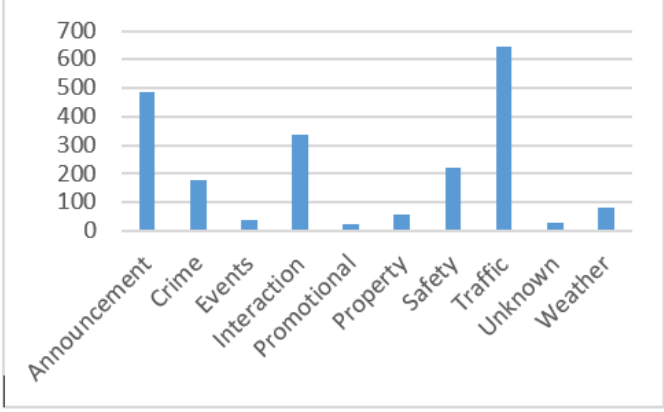

Figure 1. Tweet Count by Category

Preprocessing

The only information used for this analysis was the body of the tweet itself (the actual text of the tweet). While Twitter provides many other fields, our concern was whether we could accurately predict the classification of the tweet using the text body of the tweet.

Prior to importing the tweet into SAS, we made one change to the data because an abbreviation that means one thing in the Twittersphere means something quite different in policing – "RT". In Twitter, "RT" stands for retweet and Twitter automatically adds the term RT to the beginning of the tweet if it is being retweeted from another user. However, we found, in looking through the tweets, many occurrences of RT that were not intended to mean retweet. Consider the following tweet:

"#WPD Reports Rt 9 west at Oakland St due to a tree down http://t.co/oRfum5G8Lp"

As you can see, the "Rt" in the tweet means "Route" and not Retweet. As a result, we searched for any instance of RT that was not at the beginning of the tweet and replaced that with "route."

Following recommendations by SAS in (SAS Institute), we began our text analysis using the default settings. We used the SAS Enterprise Miner tools to stem the words, group the words into their parts of speech, and select the words that should be excluded automatically based on either their part of speech, the common stop words, or a minimum usage of the term (in this case, a term had to appear in at least 4 tweets to be considered). This allowed us to see the baseline of how well SAS could identify categories of tweets without implementing any domain specific knowledge or tweaking of the models.

Measures
To evaluate our models for effectiveness, we compute several standard scores. These include an overall model success score, as well as scores that could be used at the categorical level.

Overall Success
This is a general measure that simply measures the number of correctly classified tweets across all classifications, divided by the total number of tweets.

Precision, Recall, F-Measure
Precision, Recall, and F-Measure are all measures that we will use at the category level to evaluate how well each model is able to accurately predict tweets of a certain category.

Precision is defined as the number of true positives divided by the sum of true positives and false positives.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

For example, consider a model that has classified two tweets as related to the category Crime. This first tweet is correctly classified (True Positive) while the second tweet really should have been classified as the category Traffic (True Negative). The result would be $Precision_{Crime}$ = 1 / 1 + 1 = .5 or 50%.

Recall is the number of true positives divided by the sum of true positives and false negatives.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

For example, if we have three tweets that should have been classified as Crime but only one was, then $Recall_{Crime}$ = 1 / 1 + 2 = .33 or 33%.

F-Measure is measured using the formula 2 * Precision * Recall / Precision + Recall.

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Continuing with our prior examples $F\text{-}Measure_{Crime}$ = 2 * .5 * .33 / (.5 + .33) = .13. Each of these measures results in a positive fraction between 0 and 1 with the higher number being better.

RESULTS

To evaluate the overall success of the model, we first determined what our baseline should be. In this case, if we were to classify all tweets into a single category of "Traffic" then we would have an overall success rate of 30.86% since 645 of our 2090 tweets were of type "Traffic", which was our highest ranked category. Each of the subsequent models used the manual classification obtained from Williams et al. (2018) as the true classification for each tweet.

Our first model (Model 1_a) used the default settings in SAS. This meant that the SVD resolution was low, the Max SVD Dimensions was set to 100 (which we have called low), and the clustering algorithm used was Expectation Maximization. By default, SAS will then identify up to 40 different clusters based on the text. In model 1_a, the result was 28 different clusters.

Each cluster was assigned to one of the original 10 categories by selecting the category that occurred most frequently in that cluster (Table 1). For example, Table 1 shows that Cluster #1 comprised 101 tweets. The category "Interaction" appeared most frequently in that cluster, so that was selected as the category for that whole cluster. As a result, many clusters were classified as the same category, for example, the category "Announcement" was assigned to 13 of the 28 clusters. We then calculated the overall success rate by looking at the total number of tweets where the actual category matched the cluster category. In this model this resulted in 1,246 out of 2,090 accurately matched, or 59.62%, which is considerably higher than our baseline of 30.86%, but still only accurately predicted the category in approximately 6 out of 10 tweets (meaning our misclassification rate was 40.38%).

Table 1. Tweet Classification through Clustering

Easing Up on Dimension Reduction

Cluster	Ann.	Crime	Events	Inter.	Promo.	Prop.	Safety	Traffic	Unknown	Weather	Total
1	6			47	2		23		23		101
2	21	5	2	9	2	3	12	21		9	84
3	10	12	1								23
…											
28								68			68
Total	486	177	39	337	22	56	219	645	26	83	2090

Next we explored whether easing up on the level of dimension reduction would help with the accuracy of the clustering. Simply put, perhaps our test was complex enough that our cut-off by default was too low to detect tweets that should be similar. The two settings we used for this was to increase the maximum SVD dimensions from 100 to 200 (labeled High) and to our SVD resolution from Low to High. Both of these will result in increasing the amount of data considered when determining clusters. For our second model (1_b) we have increased the Maximum SVD dimensions to 200 (labeled High). Models 1b through 1d test each of these changes. As you can see in Table 2, the overall success in each case actually decreased. By increasing the dimensionality, we actually introduced more noise into the equation resulting in poorer performance.

Table 2. Model Comparison

Model	SVD Resolution	Max SVD Dimensions	Clustering Algorithm	# Clusters	Overall Success
Baseline	n/a	n/a	n/a	1	30.86%
1_a	Low	Low	Expectation Maximization	28	59.62%
1_b	Low	High	Expectation Maximization	13	50.48%
1_c	High	Low	Expectation Maximization	7	49.47%
1_d	High	High	Expectation Maximization	3	39.90%
2_a	Low	Low	Hierarchical	28	54.50%
2_b	Low	High	Hierarchical	7	44.31%
2_c	High	Low	Hierarchical	13	46.32%
2_d	High	High	Hierarchical	3	38.20%

Clustering Algorithm

In our next set of models, we used a hierarchical clustering algorithm. Unlike the Expectation-Maximization algorithm of the previous models, which uses a flat representation of terms, the hierarchical model creates a tree-like hierarchy to determine clusters. Models 2a through 2d use the same settings as Models 1a through 1d respectively with the difference being that the hierarchical clustering algorithm was used instead of the expectation-maximization model. As seen in Table 2, in each case, the hierarchical algorithm performed worse than the expectation maximization algorithm.

Table 3. Success Measures by Model

	Model 1_a			***Model 2_d***		
Category	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Announcement	0.521	0.617	0.565	0.296	0.899	0.445
Crime	0.855	0.367	0.514	n/a	0.000	n/a
Events	n/a	0.000	n/a	n/a	0.000	n/a
Interaction	0.388	0.415	0.401	n/a	0.000	n/a
Promotional	n/a	0.000	n/a	n/a	0.000	n/a
Property	0.571	0.429	0.490	n/a	0.000	n/a
Safety	0.402	0.470	0.434	n/a	0.000	n/a
Traffic	0.786	0.902	0.840	0.593	0.563	0.578
Unknown	n/a	0.000	n/a	n/a	0.000	n/a
Weather	0.821	0.386	0.525	n/a	0.000	n/a

Individual Classification Performance

While we now have a sense of how well various clustering techniques work across the entire set of tweets, we next examine how successful we are at the category level. In other words, how effective are we at identifying, for example, the tweets classified as Crime?

Using our 8 models we calculated Precision, Recall, and F-Measure for each category of tweet. For the purpose of comparison, we have shown the results for our best and worst overall performing models (1_a and 2_d respectively). The results are listed in Table 3. For our highest performing model there were three categories that were not identified in any of the clusters, these included "Events", "Promotional", and "Unknown". The remaining 7 categories had a precision ranging from .388 (Interaction) to .855 (Crime). This means that, for example, of those tweets classified as "Crime", 85.5% of them were accurately classified by the machine method. Recall ranged from 0 (for the three categories that did not have a corresponding cluster) to .902 for traffic. This means that 90.2% of all traffic tweets were accurately predicted.

Precision and Recall serve a different purpose and would play a different role depending on the research question. If the question was "could you find a sample of tweets related to Crime?" then, using precision, our model performs reasonably well as only about 3 in 20 of the Crime tweets would be labeled incorrectly by the machine process. However, if the question was "Can you find all of the tweets related to crime?" then, using recall, the machine process only have captured 36.7% of the overall Crime related tweets.

The F-Measure is a good overall measure used to evaluate classification performance at the category level. In model 1a the F-Measure varied from .401 to .840 for the 7 categories that could be measured. In comparison, our poorest performing model only identified 3 total clusters, two of which were both classified the same. The result was that only 2 categories were identified and neither of these performed better than model 1_a.

Domain Specific Knowledge

Now that we have documented how the various models work out of the box, we can test whether applying domain specific knowledge to the context would result in higher performing models.

As with all domains, there is policing terminology that has meaning beyond its meaning in a general setting. For example, the word "snitch" is often used as a verb (to snitch on someone to the police), but if we were reading the Harry Potter novels then snitch is a noun representing a piece of equipment in the game Quidditch.

By working back and forth with the term classification in SAS and the actual tweets, we found a series of terms that were not being used in the common context. For example, St. was being interpreted in SAS as a Title (abbreviation for Saint) while in the context of this set of tweets it represented the abbreviation for Street. To account for these differences, a series of synonyms was created so terms were being grouped with words of similar meaning. For example, the list of "rd", "road", "st", and "street" were all joined together.

Figure 2. Tweet with Domain Specific Abbreviation

Secondly, there were common abbreviations that were found such as MV for Motor Vehicle that SAS would not understand, so these abbreviations were added as synonyms for the spelled-out words. Third, there were words in the default stop-list that could contribute to the meaning of tweets, for example, titles such as Mr. and Ms. were excluded but could contribute to the "Interactions" category. Fourth, there were terms that were not in the stop-list that we felt should be added. For example, rt (meaning Retweet) was the highest occurring term in the tweets and SAS by default was using it in the clustering algorithms, we felt that for identifying categories of tweets, the fact that it was retweeted would have little importance. Finally, in an attempt to further reduce noise, the minimum number of occurrences a word had to appear in order to be included was set to 10 (or .5% of all tweets).

After making these changes, we reran the highest performing model (1_a) and the overall success rate was 55.22%, or approximately 4.4% lower than the original model (before we made these changes).

Table 4. Top Precision Rules

Target Value	Rule #	Rule	Precision	Recall	F-Measure	True Positive/Total
Traffic	14	mv accident	1	0.363	0.532	234/234
Traffic	15	Road	1	0.490	0.658	175/175
Announcement	32	daily	1	0.185	0.313	90/90
Weather	57	tornado	1	0.229	0.373	19/19
Weather	58	flood	1	0.434	0.605	17/19
Weather	59	storm	1	0.578	0.733	15/15
Crime	63	old male	1	0.266	0.420	47/47

Text Rule Builder

Finally, we tested the notion that a few singular terms might be the best classification method as suggested in (McCallum and Nigam 1998). This was accomplished by using the SAS Text Rule Builder tool. The text rule builder uses the words found in the tweets, along with the predefined category, to build rules for making classification decisions. For example, the rule might be *"Car" and "Road" and (not "Pet")* to identify a Traffic incident. The rule excluded the term "Pet" because when Pet occurred with alongside the terms "Car" and "Road" it signified that the post contained a pet safety issue, not a traffic issue. The tweets in the dataset were composed of 8,156 unique terms after applying stemming, which is a technique to reduce words to their base form .

The output from the Text Rule builder provides a series of rules along with their F-Measure allowing us to evaluate each rule individually. Using a 30% training set and a 70% testing set, the result was 107 rules, all but 2 of which were a single word (the other two were "car" & "hot" and "car" & "Woburnpolice"). 13 of the 106 rules (13.2%) had regional or local application (for example, "walthampd"). While these are not generalizable to a larger population, it does indicate that key regional and local names could be used instead of these terms.

The overall success of the rule-based model was 76.03% which is considerably higher than all other models, and means that approximately 3 out of every 4 tweets were classified correctly using the 106 rules encompassing 107 terms, or just 1.3% of all terms. There were 7 terms (singular word or multi-word term) across 4 categories that had a 100% precision rate. For example, the term "Road" is always associated with a Traffic incident while "old male" was always classified as Crime. If we were to use the singular term "Road" then we would retrieve 49.0% of all Traffic related tweets.

As far as the success of individual categories, Table 4 shows the breakdown of Precision, Recall, and F-Measure by category. Precision ranged from a low of 45.9 for the category Interaction to a high of 96.8% for Traffic. Recall ranges from 31.8% for Promotional to 92.4% for Traffic. In all categories, the F-Measure was higher in this rule-based model than in our highest performing SVD model. Recall that no SVD model even predicted the 9 primary categories (avoiding Unknown classification since these tweets could not be manually categorized with the text available). The rule-based model was effective in predicting each of the 9 categories with varying levels of success.

SUMMARY

Our research question asks which classification techniques most accurately categorize police tweets.

When using a singular value decomposition model, we found that increasing dimensions and adding complexity actually decreases success. This suggests that the tweets by police are not complex and by adding more processing power to detect weak links we instead simply added more noise to the classification challenge with no added benefit.

We find evidence that singular terms are the best predictors for police issued tweet classification and a rule-based model is well suited for this domain. Further, our findings show that a limited number of words is best at performing the classification. Our final model had just 107 terms, representing 1.3% of all terms used in the tweets, responsible for all the rules. This further substantiates our findings from the earlier models that reducing complexity leads to better results. This also helps explain why our SVD models performed worse when allowing for more complexity.

Table 5. Success Measures for Rules Based Model

	Model 4		
Category	**Precision**	**Recall**	**F-measure**
Announcement	0.830	0.562	0.670
Crime	0.891	0.74	0.809
Events	0.846	0.564	0.677
Interaction	0.459	0.875	0.602
Promotional	0.700	0.318	0.437
Property	0.683	0.732	0.707
Safety	0.832	0.726	0.776
Traffic	0.968	0.924	0.945
Unknown	n/a	0	n/a
Weather	0.956	0.783	0.861

Certain types of tweets are easier to detect than other types. Traffic and Weather are the easiest to detect while Promotion and Interaction are the hardest. Unlike the other categories, Promotion and Interaction are generic categories that do not imply a set of associations or terms to be used. We also believe that the analysis of photographs may assist us in detecting the former since many of the individual photos that we evaluated were related to self-promotion as they portrayed the police helping out in the community. We believe successfully detecting interaction could be done using Twitter metadata to tell us if the tweet is a response to another tweet, or through the use of a mask to detect the @ symbol followed by another Twitter user's username.

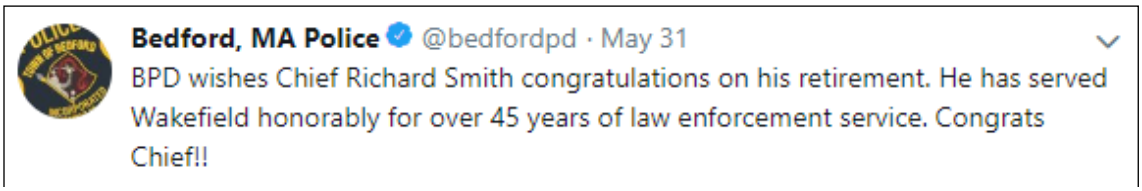

Figure 3. Promotional Tweet Difficult to Classify

The purpose of this research was to explore whether various classification techniques could be used to automatically classify tweets generated by the police department to help overcome the hurdle of manual classification thereby allowing researchers to use tweet text in their research. While the results of the clustering models were disappointing, we were encouraged by the results obtained by a simple rule-based model. What is even more encouraging is that these rules are easier to understand and explain to the general public than the more complicated clustering techniques.

This work was not intended to explore every classification technique and to present the best option, rather it was an exploration of some common techniques to evaluate where the challenges might lie in classifying police tweets. For example, with a targeted group that included only official police department accounts, we were able to avoid some of the challenges often brought up in tweet classification, namely irony and sentiment, both of which may reverse the meaning of a tweet. We avoided trying to detect sarcasm or testing for sentiment. The professional nature of these accounts led us to suspect sarcasm would not be applied. We also did not expect to see a high level of police-generated sentiment in any of the tweets, certainly not enough to be a factor in determining classification. If we were to expand this search to include those who follow the police, then these issues would need to be addressed.

By examining the results of these techniques, one can reevaluate whether the categories identified are comprehensive and individually distinct. In early tests of these models, we had distinguished between Traffic and Accident as was done in (Williams et al. 2018), but, as we mentioned in the data section, we found these topics overlapped so frequently that we were able to treat them as a singular topic. Therefore, this process can also be valuable in evaluating manual classifications.

The weakest performing category (other than Unknown) is Promotional. At closer look, we often see that self-promotion takes the form of embedding a picture of the police performing in some manner. While we ignored any photo that was part of the tweet, including the photo along with appropriate tags could serve to improve the classification of this category.

The nature of the group of tweets being considered in this analysis allowed us to avoid many of the common challenges faced when analyzing tweets. While hashtags were used, our final rule-based model suggests they did not play a key factor in determining the classification of any tweet. In browsing through the list of hashtags that were used, we noticed that these police departments appear to use hashtags on a limited basis, although this may change over time.

LIMITATIONS AND FUTURE WORK

There are several limitations with this research. From the practice perspective, the police departments were limited to just 5 different departments in a single U.S. state. There may be categories and topics of interest outside of these departments that the rule-based model would miss. Additional study is needed to suggest best practices or social media policies based on the results herein.

From a methodological perspective, the clustering models relied on selecting the category that occurred most frequently in that cluster. Clearly, if a new set of tweets had not been manually classified, this would not be possible. Instead, analysis would have to rely on key terms from each cluster, which may not match the highest occurring category. Therefore, our figures for models 1a through 2d are the maximum possible values based on this technique.

Additionally, also related to our clustering models, we allowed SAS to select the ideal number of clusters in each model yet we did not penalize models that had higher cluster counts. Taken to the extreme, we could have achieved 100% success by simply creating 2,090 clusters, each with a singular term. To evaluate these models sufficiently, some penalty should be applied for each increase in cluster count.

Finally, our assignments of categories for the SVD models was based on the most frequently occurring category in the results. However, as you can see in Figure 1, not all categories had an equal number of tweets. This likely caused several categories never to be assigned since the overall tweet count for those categories was significantly lower than other categories. So, while our approach resulted in the highest possible overall success, it did so while penalizing small categories. An equal weighting scheme could be applied if the desired result was to identify specific tweets in categories with fewer tweets.

To implement the rule-based model across a wider sample of police departments, a larger sample of pre-classified tweets is needed to generate the rules. This would overcome localized limitations and potentially cause new categories to emerge. In addition, we recommend that consideration should be taken as to whether certain masks could be used in place of local terminology (ex. City name +"pd" to overcome the local "walthampd").

We have shown that an effective rule-based model can be used to classify tweets across similar posting sources (i.e., police departments). This method would allow researchers to let the data speak for themselves and avoid some challenges associated with survey data (e.g., perceptions, low response rate, costs). We encourage others to build upon our findings.

REFERENCES

Edlins, M., & Brainard, L.A. (2016). Pursuing the Promises of Social Media? Changes in Adoption and Usage of Social Media by the Top 10 Us Police Departments. *Information Polity, 21*(2), 171-188.

Fernandez, M., Dickinson, T., & Alani, H. (2017). An Analysis of Uk Policing Engagement Via Social Media. *International Conference on Social Informatics*: Springer, 289-304.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*: Association for Computational Linguistics, pp. 581-586.

International Association of Chiefs of Police. (2016). 2016 Law Enforcement Use of Social Media Survey.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the Omg!, *Fifth International AAAI conference on weblogs and social media*.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 workshop on learning for text categorization*: Citeseer, pp. 41-48.

Sanders, A., & DeVault, C. (2004). Using Sas® at Sas: The Mining of Sas Technical Support, *SUGI 29*).

SAS Institute. *Sas Book on Text Mining and Analysis*.

Sudhan, H.H., Garla, S., & Chakraborty, G. (2012). Analyzing Sentiments in Tweets About Wal-Mart's Gender Discrimination Lawsuit Verdict Using Sas® Text Miner. *SAS Global Forum*.

Tran, M. (2014). #Mynypd Twitter Callout Backfires for New York Police Department.

U.S. Department of Justice (2015). Local Police Departments, 2013: Personnel, Policies, and Practices.

Williams, C. B., Fedorowicz, J., Gulati, J., Haughton, D., Xu, J., Kavanaugh, A., Mentzer, K., Sawyer, S., & Thatcher, J.B. (2015). Leveraging Social Media: The Community Policing Case. *3rd Annual Research Showcase*. Waltham, MA: Bentley University.

Williams, C.B., Fedorowicz, J., Kavanaugh, A., Mentzer, K., Thatcher, J.B., & Xu, J. (2018). Leveraging Social Media to Achieve a Community Policing Agenda. *Government Information Quarterly, 35*(2), 210-222.

Zubiaga, A., Spina, D., Martínez, R., & Fresno, V. (2015). Real-Time Classification of Twitter Trends. *Journal of the Association for Information Science and Technology, 66*(3), 462-473.