Improving Classification of Twitter Behavior During Hurricane Events

Kevin Stowe, Jennings Anderson, Martha Palmer, Leysia Palen, Ken Anderson
University of Colorado, Boulder, CO 80309
[kest1439, jennings.anderson, mpalmer, palen, kena]@colorado.edu

Abstract

A large amount of social media data is generated during natural disasters, and identifying the relevant portions of this data is critical for researchers attempting to understand human behavior, the effects of information sources, and preparatory actions undertaken during these events. In order to classify human behavior during hazard events, we employ machine learning for two tasks: identifying hurricane related tweets and classifying user evacuation behavior during hurricanes. We show that feature-based and deep learning methods provide different benefits for tweet classification, and ensemble-based methods using linguistic, temporal, and geospatial features can effectively classify user behavior.

1 Introduction

Identifying relevant information for natural disaster and other hazards is a difficult task, particularly in social media, which is often noisy. Understanding people's behavior during events is an important task for both researchers studying human responses to hazards after events and real-time processing of disaster-related information. Keyword searches can be an effective first pass, but are insufficient to fully understand user behavior and can generate large numbers of both false positives and false negatives. To improve our ability to study behavior during crisis events, we employ supervised machine learning for two tasks: identifying tweets that are relevant to hurricane events and classifying Twitter users' evacuation behavior.

2 Task One: Improving Tweet Classification

Twitter data is often difficult to understand due to limited length of tweets and the noise inherent in the medium. As a result, there is a variety of research in attempting to effectively identify and classify tweets. There are multiple studies in classification of flu-related tweets (Culotta, 2010; Aramaki et al., 2011). One relevance classification approach is Lamb et al. (2013), which initially classifies tweets for relevance and then applies finer-grained classifiers. They build classifiers using syntactic and Twitter-specific features to detect awareness versus infection, self versus others, and whether tweets are relevant to the flu or not.

Sriram et al. (2010) propose a somewhat more specific system, classifying tweets into general categories like news, events, and opinions, achieving accuracies between .85 and .95 depending on category. Sankaranarayanan et al. (2009) perform a similar task, classifying tweets into either news or non-news. Recently, the work of Volkova et al. (2017) attempts to classify suspicious and trusted tweets. They find that deep learning models outperform feature-based models, but linguistics features can be helpful. They report F1 scores of between .88 and .92 depending on the category classified.

For our first task of relevant tweet classification, we employ supervised machine learning to predict whether individual tweets are relevant to a hurricane. This study focuses on the Hurricane Sandy event in October of 2012. This hurricane made landfall on the eastern seaboard of the United States on October 29, causing massive damage to many areas including New York and New Jersey. To collect data for this event, we initially performed a collection capturing all tweets

Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media , pages 67–75,
Melbourne, Australia, July 20, 2018.

Tweet	Relevant
For the love of that money.....	n
Lol the struggle for gas and Power 	y
where u been hiding at through this storm	y
Smh I still don't get to play Halo 4 yet...	n

Table 1: Sample Tweet Classification Stream

using the following keywords:

DSNY, cleanup, debris, frankenstorm, garbage, hurricane, hurricanesandy, lbi, occupysandy, perfectstorm, sandy, sandycam, stormporn, superstorm

This generated approximately 22.2 million unique tweets from 8 million users. We then identified users who had geo-tagged tweets within areas that were heavily impacted by the event. This allowed us capture users who were likely to be significantly impacted and local to the event. From these we randomly selected 105 users, collecting the tweets from a week before landfall to a week after, resulting in 25,474 tweets. We annotated these tweets for hurricane relevance (two annotators, agreement approximately .9). Our task is to classify for each user which tweets are relevant (Table 1).

We developed a standard feature-based machine learning classifier and compare it to several deep learning approaches. We split our data into training (60%), validation (20%), and test (20%) sets, tuning each model on the validation set and evaluating on the test data.

2.1 Feature-based

As a baseline for feature-based classification, we follow the setup and features of Stowe et al. (2016), who employ support vector machines and linguistic features to classify hurricane related tweets. As a baseline, we re-implement this approach, leaving out features that appeared to have negligible contribution. We used the following features from their set:

- Bag of words based on Pointwise Mutual Information (PMI) for unigrams, bigrams, and trigrams. We chose the n terms with highest PMI for positive and negative classes, with n set to 200 as was determined in validation. Selecting the bag of words lexicon based on PMI significantly improves results over using the full set of words.

Model	F1	Prec	Recall
Stowe et al (SVM Baseline)	.769	.886	.678
Multi-layer Perceptron	**.834**	.886	.788
Convolutional NN	.815	.874	.763

Table 2: Tweet Classification Results

- The time of the target tweet, using a one-hot vector representing the time bin of the target tweet. Through validation we chose to use 384 bins, or one per hour.
- Average word embeddings for each tweet. We experimented with using Google News vectors generated using word2vec (Mikolov et al., 2013) and Glove Twitter embeddings (Pennington et al., 2014). We selected the Google News vectors, as they had the best performance.

For comparison, we employ two deep learning approaches: a multi-layered perceptron (MLP) and a convolutional neural network (CNN).

2.2 Multi-layered Perceptron (MLP)

For our MLP, we started with inputing each tweet as a collection of words, padded up to length 25. We used an embedding layer of dimension 300 using the pretrained Google News vectors, and fed this through a 50 node dense layer using a rectified linear unit (relu) activation with a dropout rate of .5. This was then fed into the output layer, using sigmoid activation to predict either relevant or irrelevant. The model was trained using categorical hinge loss, running 50 epochs.

2.3 Convolutional Neural Network (CNN)

Convolutional neural networks incorporate local word context using convolutions of words within a contextual window, and have proven effective in a variety of sentence classification tasks (Kim, 2014; Li et al., 2017). As tweets can be considered a sentence, we experiment with using CNNs for relevance classification.

We follow the approach of Kim (2014), using an embedding layer (from the Google News vectors), which is then fed into a convolutional layer. We use kernel sizes of 2, 3, and 4, with 16 filters per kernel size. We use max pooling to combine the outputs, with a pool size of 4. Finally, we use a fully connected layer to the binary output nodes, using sigmoid activation to predict relevance.

Both deep learning models improve over the re-implemented SVM baseline. However, the CNN

doesn't improve over the basic multi-layer perceptron.

2.4 Effects of Context

Sentence classification is a common task, and it has been applied effectively to tweets. However, most classification for Twitter data is done on individual tweets, without regard to their larger context. This causes an impoverished information environment: knowing the context a tweet is present in from a user's perspective provides valuable information about the meaning of the tweet.

Because of this, we experimented with using contextual models to predict tweet relevance. We experimented with using the same SVM model above, experimenting with expanding the feature window to include more context, as well as adding additional contextual tweets the MLP model. In both cases, we used contextual windows from 1-16 words before and after the tweets. We found that performance decreased consistently as more context was added, and using only the target tweet yielded the best results

We also experimented with using sequence taggers, specifically a long short-term memory (LSTM) network. We input each user as a training batch, treating the tweets they produced chronologically as a sequence. Our results using the LSTM model were much lower than the non-sequence taggers (.65 compared to .83). Tuning model size and dropout, as well as adding bidirectional and attention layers failed to significantly improve performance.

From the data, it appears that context is vital for determining tweet relevance, but our models have not been able to capture the significance. We believe this is due to the irregular nature of helpful context. In tweet streams, it is often the case that one particular tweet in the context is necessary to understand the target, but the location in context of the tweet is not consistent. Because of this inconsistency, the model cannot reliable determine which element in context is contributing the necessary information. As a future goal, we aim to incorporate better methods of representing context that can filter out contextual tweets that likely don't influence the target.

2.5 Effects of Data Size

As each event is unique and other kinds of natural hazards are likely to pose completely new problems, we would ideally like to be able to generate new classifiers with as little data as possible. We experiment with varying the size of our training data to assess how much is necessary to reach peak performance. We held out 20% of our data as a test set, and then trained classifiers incrementally, adding 100 instances of training data at a time. We also tested the effectiveness of combining models by implementing a combined classifier. This classifier uses the output of the MLP and the SVM as features for training a logistic regression classifier. The results of these classifiers as training data is added are shown in Figure 1.

The SVM achieves strong recall very quickly, at over .8 with only 5,000 training instances. The perceptron follows an opposite pattern, with precision over .9 at 5,000 but very low recall. The SVM is consistently improving at around 2,500 training instances, and shows only minimal improvement after 7,5000. The perceptron is much more irregular, being ineffective until nearly 7,500 instances and leveling off near 12,500.

We believe that precision is more important for this task, as there are such a large number of tweets available, it is more important to identify tweets correctly than to capture all of them. However, the perceptron takes more data to be consistent. Combining classifiers in this case doesn't improve performance over either the SVM or MLP individually, although the logistic regression approach is comparable. The best approach for extending classification novel events is to assess whether precision or recall is more important, and select the individual classifier that fits the goals of the research.

Classification of tweets can be improved by employing deep learning models, which significantly outperforms feature-based methods. Comparisons to other work are difficult to the differences in tasks. We do not achieve the F1 scores of Volkova et al. (2017) or Sriram et al. (2010), both between .85 and .95, but the tasks are likely too different for meaningful comparison.

3 Task 2: Evacuation Classification

Tweet classification provides information about user behavior as users tweet about their experiences and actions as the events are unfolding. At a broader level, we can also use tweet streams from a user to attempt to determine their evacuation behavior during an event. For this, we need to examine their entire stream and understand both their

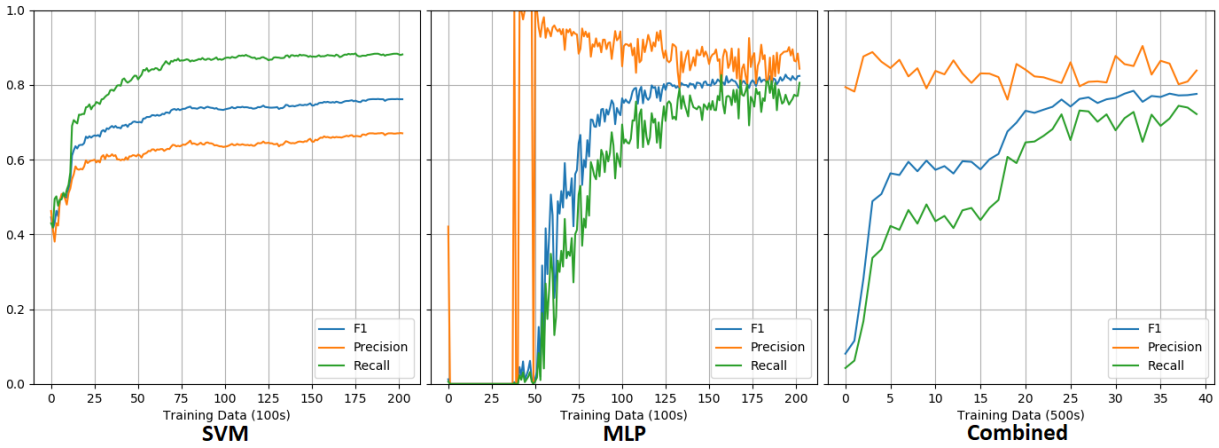

Figure 1: Classification as Training Data is Added

language and their actions. In this section we describe our annotation of Twitter users' evacuation behavior, and show that linguistic and geospatial data can used for classification.

User level classification for Twitter users is a well known problem for many domains. One common task is identifying political affiliation. Both linguistic and non-linguistic features have proven effective in classifying political leanings of users in Twitter data (Tatman et al., 2017). Preoiuc-Pietro et al. (2017) provide a method for identifying whether users are liberal or conservative, and point to a variety of user level classifications that can be predictive of political ideology. These user-level attributes apply generally; we intend to classify users based on a particular behavior they engage in (evacuation or sheltering in place).

More similar is Sanagavarapu et al. (2017), who predict whether users participate in events that they are tweeting about. They use linguistics features coupled with support vector machines to predict users' participation in specific events, which parallels our task of predicting a user's event-related behavior.

In the domain of crisis informatics, recent work by Martín et al. (2017) identifies evacuation patterns, using aggregates of geo-located tweets as well as particular user behaviors. However, they don't empirically validate their observations, and thus don't attempt statistic learning for classification. Another study from Yang et al. (2017) studies user behavior during crisis events, using linguistic and spatial features to analyze shifting sentiment during Hurricane Sandy. While they focus on keyword tweets clustered geographically, they show that geospatial features are helpful for analysis of user attitudes during crises.

3.1 Data

Our analysis is focused on users that are potentially at risk, but these users are difficult to identify due to the noisiness of Twitter data. To alleviate this problem, we attempt to identify vulnerable users using geospatial information. For our data, location-enabled tweets include any tweet returned by the Twitter API with a precise point-location attribute. This is sometimes the precise latitude and longitude of the user's mobile device; however, and more common in recent years, these are more general locations that, while encoded in the tweet as a single geographic coordinate, represent businesses or more general regional locations. These often include cross-posts from other social media services that track location such as FourSquare, Swarm, or Instagram. Examples of these locations include: "Starbucks" (as an exact store) or "South Beach" (as a region).

For Hurricane Sandy, we used bounding boxes for Evacuation Zone A in New York City as well as boundaries of the coastal counties of New Jersey to define geographically vulnerable areas. Each of these areas were under mandatory evacuation orders and generally exhibited high levels of geographic risk to the storm.

3.2 Spatial Clustering

To reduce the noise and identify the most important locations for a user, we apply a clustering algorithm to all of the tweets for a given user. We use Density Based Spatial Clustering (DBScan) to cluster each user's tweets based on their coordinates (Ester et al., 1996). We chose this algorithm

for two reasons. First, it does not require that we declare a particular number of clusters ahead of time. Since we cannot make any assumptions about a user's consistent located-enabled Twitter activity, we do not know how many clusters will best represent the recurring locations for any given user. Second, it does not require that all points be classified. This allows for rigid, similar sized clusters with separate unclassifiable points. Through empirical analysis of our data, this is critical to understanding a user's recurring tweet locations because users tend to tweet very irregularly (spatially speaking): on a moving bus or train, for example.

Once these spatially outlying points are marked as noise, we focus analysis on locations of consistent, recurring Twitter behavior, such as one's residence or workplace. Our clustering parameters require a user to have at least five tweets within 100 meters of one-another within the three-month period of study. These parameters are stricter than those used in Jurdak et al. (2015) and were decided through empirical analysis of spatial tweet distributions of a few users. Since the purpose of the clusters is to identify areas of work or residence that may be at risk of a coastal hazard, 100 meters allows the clustering to account for some noise and inaccuracies in the reported location over the entire study period. We remove any users who do not have at least one identifiable cluster.

3.3 Temporal Clustering

To learn about a user's regular (non-storm) Twitter behavior, we identify their temporal tweeting patterns up to the time of the storm. To generalize this over the entire period of study, we look specifically at times of tweets per week. Given the regular diurnal Twitter activity among users, we next cluster the tweets by time of day and day of week to establish a weekly tweeting distribution for each user. Krumm et al. (2013) use a similar method of discerning home locations based on the time one is active, based on the American Time Use Survey. First, we distinguish days as weekdays or weekends and then split these days into six four-hour periods. The resulting 12 time bins distinguish between common home and working hours. Of these times, weekday evenings generally see the most Twitter activity.

3.4 Spatio-Temporal Clustering: Home Locations

Co-occurrences between the geo- and temporal-clusters identify likely home clusters as distinct from work or school clusters. For example, if a user's tweets from *geo-cluster A* occur primarily during weekdays from 12-4pm while *geo-cluster B* primarily includes tweets from weekday evenings from 8pm-12am, then we may infer that *geo-cluster A* could represent that user's school or workplace while *cluster B* could represent their home. To perform this identification of a user's before-storm home location, we then identify geo-clusters that commonly co-occur with the following specific time bins that represent *home times*: Weekdays between 12-4am, 4-8am, and 8pm-12am. The geometric centroid of the cluster with the most tweets during these times is said to be the user's home location.

Note that these home locations don't necessarily represent where the user lives. While we qualitatively observe that these home locations usually appear to be correct, they also can be gyms, offices, and other places that the user typically tweets from. We'll see in section 3.6 that home location information is a good predictor of evacuation behavior, regardless of whether it represents an actual 'home' or merely a location of consistent behavior for a particular user when daily life is not interrupted by a major storm. If this location lands within the geographically vulnerable areas under mandatory evacuation described above, this user is said to geographically vulnerable. Furthermore, the empirically observed accuracy of this approach to determining a user's home location invites further research that optimizes the clustering (both spatially and the temporal bins) to improve detection of a user's home-location based on their geo-located social media activity.

The simplicity of our approach combined with the observed accuracy suggests that users are likely not aware of the extent and accuracy of the public geo-trace they are producing through their social media activity. All of the tweets used for this work were posted to the user's Twitter timeline for public consumption. As a first step to protecting user's privacy, we do not publish the user's Twitter handle, against the formal guidelines for republication of Twitter data. Further, we intentionally do not show a larger-scale rendering of their calculated home location. For these reasons,

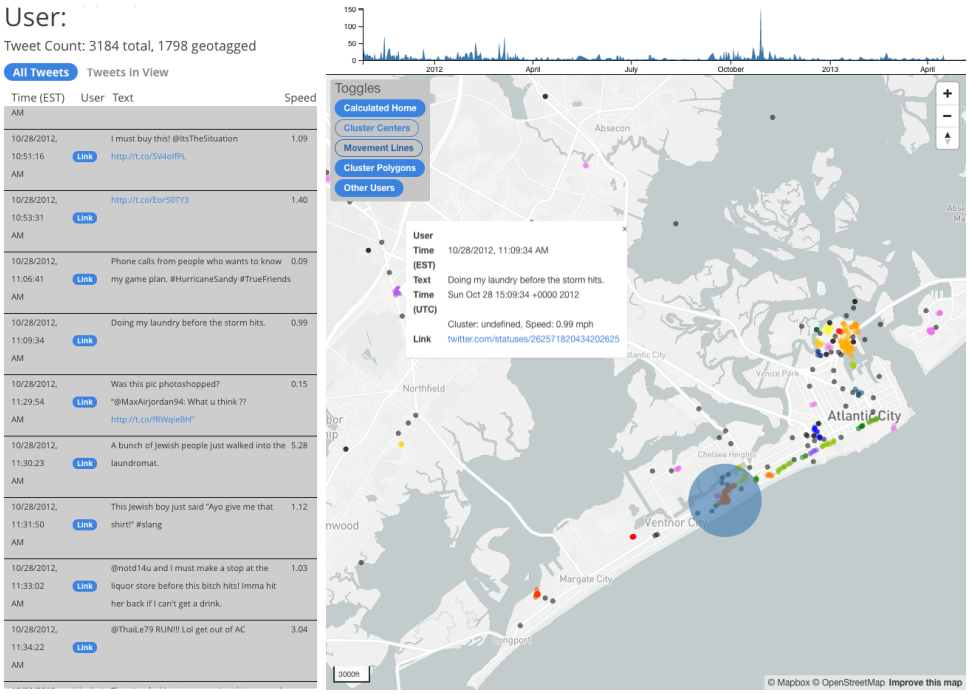

Figure 2: A screen shot of our annotation tool. The timeline across the top shows this user was most active at time of landfall. Tweets are displayed in chronological order on the left and an interactive map on the right shows tweets, colored by different clusters. The transparent blue circle indicates the calculated home location. Just before landfall and throughout the duration of the storm, the user is tweeting from a different location (see popup) further inland than their calculated home: a strong signal of evacuation.

and in part because the data is now over five years old, we choose to publish the data as-is, without their identifiable Twitter handle. These are self-imposed ethical responsibilities because as far as the data providers are concerned, this is public data. We hope this situation invites further conversation around social media privacy, information sharing, and more formalized ethical standards in social media research concerning these highly personalized data traces.

3.5 Annotation

Our perception of spatially derived evacuation patterns is clear: geographically vulnerable users tweet from their vulnerable locations before the storm and then do not tweet from this location at landfall. However, few users have such clear cut movement profiles. Furthermore, programmatically searching for this behavior yields a troubling amount of false positives. Just because a user is not tweeting from home does not mean they have chosen to evacuate. These complex user behaviors led us to develop a tool and annotation process for determining individuals responses to the events.

Our annotation involves determining if a user evacuated, sheltered in place, or their behavior was unclear based on the available data. Each user was given one of these categories based on both their tweet content and movement patterns as inferred by manual inspection. This involved developing a framework for displaying tweets on a map over a sliding window of time, allowing annotators to easily identify what users were saying at which locations, thus giving the capability to determine possible evacuation behavior quickly and accurately. Using this tool, we tagged 200 users with evacuation, sheltering in place, or unclear, along with a confidence score for evacuation and sheltering in place.

Note that this annotation process has inherent problems: we can only indicate whether we believe a user evacuated based on their tweets and geo-location. We can not prove that any user evacuated based only on these limited resources. So while annotators tend to agree on whether they believe a person took a particular action (κ=.705 for tweets annotators were confident of the correct answer), the analysis is not objectively verified.

Tweets	Coords	Time
Hurricane Party!	40.6,-73.9	12/29 14:03
I had to evacuate this is bull	40.8,-72.9	12/29 19:26
East NY 4ever!	40.8,-73.4	12/30 11:45
Prediction	Evacuated	

Table 3: Sample Evacuation Prediction

3.6 Classification

We employ supervised machine learning to predict each user's possible actions during each event. This is done by employing word embeddings to represent tweet semantics combined with temporal and spatial features generated from tweet metadata. We treat each user's full contextual stream (all the tweets they produced from a week before to a week after landfall) as a document (see Table 3 for an example). As a baseline for classification, we start with the average embedding over all the words in the contextual stream, providing a simple document-level embedding.

3.6.1 Temporal Information

We've seen from section 3.1 that users' tweeting behavior varies greatly based on time. In order to capture this, we split each user into a series of time bins. For each bin, we generate the average embedding over all the tweets in the time slice. These embeddings are concatenated and supplied as input to the classifier. We experimented with a variety of bin sizes from 4 hours to 4 days. Smaller bins capture more specific data, but are often contain too few tweets to be useful. Larger bins provide more consistent, general information.

3.6.2 Spatial Information

We combine information from geo-tags with word embeddings to generate more accurate representations of user behavior. For each temporal bin generated above, we calculate a handful of spatial features. First, we calculate the average location of the user during that bin, using the mean latitude and longitude of each tweet in that bin that contains a geo-tag. We then use this to determine the geometric distance from the average location in that bin to the calculated home location from section 3.4. This is a simple scalar feature indicating their distance from their typical home location. As a second spatial feature, we calculated the average distance of each tweet within a bin from the starting location of that bin, which indicates the average amount the user moved during that time.

3.6.3 Relevance Filtering

In most cases the majority of tweets a user produces are irrelevant to a particular event. This creates additional noise in each time bin, making it hard to predict behavior. We employ the relevance classifier above, trained on the full dataset, and use it to predict relevance for each user's tweets. We then restrict the features above to only tweets that the classifier deemed relevant.

We evaluate Logistic Regression, Support Vector Machines, and Naive Bayes algorithms for user classification. We experimented with deep learning methods, but they showed much lower results, perhaps due to the small size of the dataset. Support vector machines provided the best performance on the baseline, and was used to evaluate additional features. We performed 10-fold cross validation over users. Table 4 shows the results of adding each feature type and bin size. Each column represents the size of the temporal bin used. The "All" column uses only one bin, with all the user's tweets averaged. In this case the Distance from Home Location is the distance from their overall average location to the location of their calculated home location from section 3.4.

Bin sizes from 1 to 4 days are most effective, with distance from home location being the best feature. Note that we did not objectively verify these home locations: the classifier uses this feature effectively regardless of whether it represents the user's real home, or just a location they regularly tweet from. Relevance filtering does not provide consistent improvement, which may be due to data sparsity. Any filtering reduces the amount of tweets available for each bin, making the classification task more difficult.

These four features (word embeddings, temporal and spatial information, and relevance filtering) all provide different ways of understanding user behavior. Because they represent the data in different ways, they are capable of classifying different sections of the data accurately. We leverage this by employing ensemble classification employing these features.

3.6.4 Ensemble Classification

To combine feature's benefits, we use each of the 48 classifiers generated for Table 4. We combine these classifiers incrementally, starting with the classifiers that had the best performance in cross validation. We trained each classifier on 50% of the data and evaluated it on the remaining 50%.

Feature	4 hours	8 hours	1 day	2 days	4 days	All Days
Word Embedding Average	.467	.491	.529	.505	.555	*.529*
+Relevance Filter	.481	.484	.489	**.606**	.526	.533
Distance from Home Location	.610	.692	**.695**	.686	.639	.674
+Relevance Filter	.657	.621	.672	.661	.608	.664
Average Movement per Time Bin	.484	.506	.504	.587	**.641**	.533
+Relevance Filter	.541	.501	.526	.566	.574	.508
All	.533	.504	**.550**	.524	**.550**	.531
+Relevance Filter	.484	.485	.493	.486	.534	.534

Table 4: Classification Results (F1) for Each Feature Type and Bin Size. **Bold** indicates the best result for that feature, *italics* is our word embedding baseline.

We then weighted each classifier's classification by the F1 score it received in cross validation on the training set. This allowed for more classifiers to be added providing additional information, but still favoring the classifiers that performed best in training. Results of the incremental addition of classifiers are shown in Figure 3.

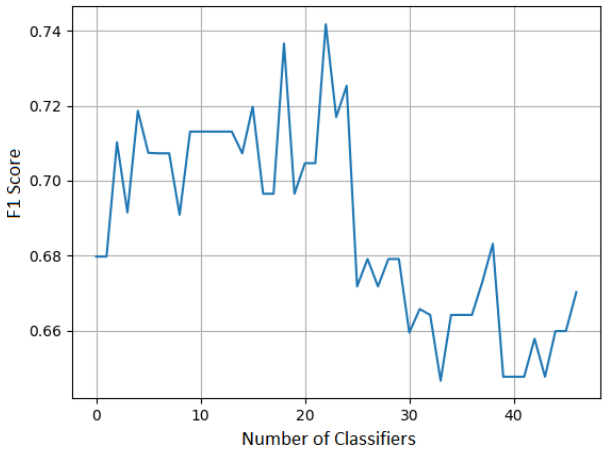

Figure 3: System performance (F1) as classifiers are added.

3.7 Analysis

Adding additional classifiers improves performance initially but after a certain point the added classifiers decrease performance. The best performance is achieved using around 20 classifiers, which include those trained on all three individual features for a wide variety of time bins. While the first 16 classifiers are based on distance from home cluster, the addition of word embedding- and movement-based classifiers can yield improved performance. As more classifiers are added, performance drops, likely because when the less accurate classifiers are added they degrade performance.

Performance on user classification varies greatly depending on the classification method and windows used. The basic word embedding baseline over all tweets performs poorly (.529). Prediction based on distance from a user's home location using a bin size of 1 day is the best single classifier (.695), and distance from their calculated home performs best across all bin sizes. The best F1 achieved through ensemble methods is .741, a considerable improvement over the performance of the best individual classifier .694.

While it is difficult to compare these results to previous work, as the task has not yet been attempted, there are some relevant comparisons. Sanagavarapu et al. (2017) report classifying users' participation in events with F1 scores varying from .52 to .74. They show that different features yield different results based on the time period, which parallels our results.

4 Conclusions

Evacuation behavior is difficult to predict, but can be done by leveraging both linguistic and geospatial features. More data and better representations of movement could improve this classification, but the changing nature of Twitter use is making precise geospatial data increasingly rare and harder to make us of for behavior classification in this medium.

Our relevance classifier achieves an F1 score of near .83, and needs refinement to be effectively employed in this domain. Further improvements to classification can be made be effectively incorporating tweet context. In order to make use of this classification, we intend to experiment with real-time relevance classification, which will allow us to better understand user behavior live as events unfold.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1576.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231.

Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. Understanding human mobility from Twitter. *PLoS ONE*, 10(7):e0131469.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

John Krumm, Rich Caruana, and Scott Counts. 2013. Learning likely locations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7899 LNCS, pages 64–76.

Alex Lamb, Michael J Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795.

Shen Li, Zhe Zhao, Tao Liu, Renfen Hu, and Xiaoyong Du. 2017. Initializing convolutional filters with semantic features for text classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1884–1889, Copenhagen, Denmark. Association for Computational Linguistics.

Yago Martín, Zhenlong Li, and Susan L. Cutter. 2017. Leveraging twitter to gauge evacuation compliance: Spatiotemporal analysis of hurricane matthew. *PLOS ONE*, 12(7):1–22.

Tomas Mikolov, Kai Chen, Greg Corrado, and Dean Jeffrey. 2013. Efficient estimation of word representations in vector space.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Daniel Preoiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Krishna Chaitanya Sanagavarapu, Alakananda Vempala, and Eduardo Blanco. 2017. Determining whether and when people participate in the events they tweet about. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 641–646, Vancouver, Canada. Association for Computational Linguistics.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, and Jon Sperling. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.

Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. 2016. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6, Austin, TX, USA. Association for Computational Linguistics.

Rachael Tatman, Leo Stewart, Amandalynne Paullada, and Emma Spiro. 2017. Non-lexical features encode political affiliation on twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 63–67, Vancouver, Canada. Association for Computational Linguistics.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Min Yang, Jincheng Mei, Heng Ji, zhao wei, Zhou Zhao, and Xiaojun Chen. 2017. Identifying and tracking sentiments and topics from social media texts during natural disasters. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 527–533, Copenhagen, Denmark. Association for Computational Linguistics.