Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol

Robert Munro
Department of Linguistics
Stanford University
Stanford, CA, 94305
rmunro@stanford.edu

Abstract

Crisis-affected populations are often able to maintain digital communications but in a sudden-onset crisis any aid organizations will have the least free resources to process such communications. Information that aid agencies can actually act on, 'actionable' information, will be sparse so there is great potential to (semi)automatically identify actionable communications. However, there are hurdles as the languages spoken will often be under-resourced, have orthographic variation, and the precise definition of 'actionable' will be response-specific and evolving. We present a novel system that addresses this, drawing on 40,000 emergency text messages sent in Haiti following the January 12, 2010 earthquake, predominantly in Haitian Kreyol. We show that keyword/ngram-based models using streaming MaxEnt achieve up to F=0.21 accuracy. Further, we find current state-of-the-art subword models increase this substantially to F=0.33 accuracy, while modeling the spatial, temporal, topic and source contexts of the messages can increase this to a very accurate F=0.86 over direct text messages and F=0.90-0.97 over social media, making it a viable strategy for message prioritization.

1 Introduction

The recent proliferation of cellphone technologies has resulted in rapid increases in the volume of information coming out of crisis-affected regions. In the wake of the January 12, 2010 earthquake in Haiti local emergency response services were inoperable but 70-80% of cell-towers were quickly restored. With 83%/67% of men/women possessing a cellphone, the nation remained largely connected. People within Haiti were texting, calling and interacting with social media, but aid agencies were not equipped to process so much information. This is because in any sudden onset crisis, this flood of information out of the region coincides with crisis-response organizations already working at capacity as they move in. The problem definition is clear: *how can we filter this information to identify actionable intelligence to support relief efforts?*

The solution is complicated. There may be few resources for the language(s) of the crisis-affected population and the ratio of actionable to non-actionable information will often be very large, especially when reported through social-media and other non-official channels. In the absence of existing electronic resources models must be built on-the-fly and account for substantial spelling variations. The definition of what constitutes actionable intelligence will often be context-specific and changing, too. In the data used here 'actionable' changed quickly from Search and Rescue to any medical emergency and then to include clusters of requests for food and water. The models will therefore need to be time-sensitive or otherwise adaptive.

The system presented here attempts to address these problems, finding that the accurate identification of actionable information can be achieved with subword models, the automatic identification of topics (categories), and spatiotemporal clustering, all within a streaming architecture. It is evaluated using 40,811 emergency text messages sent in Haiti following the January 12, 2010 earthquake.

2 Evaluation data

Three sets of short-messages are used, all from between January 12 and May 12, 2010:

1. *Mission 4636*. 40,811 text-messages sent to a free number, '4636', in Haiti.
2. *Radio Station*. 7,528 text-messages sent to a Haitian radio station.
3. *Twitter*. 63,195 Haiti-related tweets.

The *Mission 4636* messages were translated, geolocated and categorized by a volunteer online crowdsourced workforce, predominantly from the Haitian diaspora, and by paid workers within Haiti (Munro, 2010). English-speaking crisis-mappers identified actionable messages and refined the coordinates and categories. The categories are a standard set of UN-defined emergency categories with some additions (48 total). The definition of an 'actionable' message was defined by the main responders to the messages, the US Coast Guard and the US Marines working with Southern Command, and included any messages with an identifiable location that contained a request for medical assistance, Search and Rescue, water-shortages, clusters of requests for food in areas not known to aid workers, security, and reports of unaccompanied minors.

The radio station and Twitter messages are not structured or geolocated. They are used here as potential false-positives in a needle-in-a-haystack scenario where the majority of messages are irrelevant. A recent Red Cross survey (2010) found that nearly half the respondents would use social media to report emergencies, so this is a realistic scenario.

3 Streaming models

Supervised streaming-models attempt to classify an incoming stream of unlabeled items by building up training-data on past items (Aggarwal, 2006; Hulten et al., 2001; Babcock et al., 2002). The models are often time-sensitive with training data either weighted towards or exclusively consisting of more recently seem items, especially in the context of concept drift or bounded memory (Zhang et al., 2008).

There is a penalty for applying GOLD labels to past items for training: either only a subset can be labeled or there is a time-delay. When only a subset can be labeled the scenario is similar to that of *active learning* (Cohn et al., 1996; Tong and Koller, 2002). When there is a delay not all past items are immediately available, meaning that short-term concept drifts might not be adapted to. In both cases, accuracy is continually evaluated over incoming items.

Here, the time-delay penalty was used for all Mission 4636 messages as it is closer to the actual scenario where each potential emergency message is ultimately read by a person but with potential delays from sudden bursts and backlogs.

The messages were divided into 100 temporally ordered sets. Each set belongs to one epoch i in the streaming architecture, R, such that R_i is evaluated on $R_0, \ldots, R_{i-1}$ (R_1 is evaluated on a model trained on R_0; R_2 is evaluated on a model trained on $\{R_0, R_1\}$; R_3 is evaluated on a model trained on $\{R_0, R_1, R_2\}$, etc.). The accuracy is therefore calculated over $R_1, \ldots, R_{99}$ (all but the first set).

The results here report a system using a MaxEnt learning algorithm with Quasi-Newton optimization and a convergence tolerance of 10^{-4}. Changing parameter settings or swapping out the learning algorithm with linear and quadratic kernel SVMs made little difference in accuracy (see discussion of other models below).

4 Features (F)

G : Words and ngrams.
W : Subword patterns (extended definition below).
P : Source of the message (phone number).
T : Time received.
C : Categories ($c_{0,\ldots,47}$).
L : Location (longitude and latitude).
$L_\exists$: Has-location (there is an identifiable location within the message).

G, W, P and T were calculated directly from the messages. C and $L_\exists$ were predicted using independent streaming models and L was clustered through tiling (see below).

4.1 Hierarchical streaming models for has-location ($L_\exists$) and category (C)

The SMS protocol does not encode locations or categories. The streaming model was extended to a two-level hierarchical architecture so that we could use (predicted) locations and categories as features.

Nou tigwav,nou pa gen manje nou pa gen kay. m.			
‘We are Petit Goave, we don’t have food, we don’t have a house. Thanks.			
Actionable	-72.86537, 18.43264	1/22/2010 16:59	2a. Food Shortage, 2b. Water shortage
Lopital Sacre-Coeur ki nan vil Milot, 14 km nan sid vil Okap, pre pou li resevwa moun malad e l’ap mande pou moun ki malad yo ale la.			
‘Sacre-Coeur Hospital which located in this village Milot 14 km south of Oakp is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.’			
Actionable	-72.21272, 19.60869	2/13/2010 22:33	4a. Health services
Mwen se [FIRST NAME] [LAST NAME] depi jeremi mwen ta renmen jwenm travay.			
‘My name is [FIRST NAME] [LAST NAME], I’m in Jeremi and I would like to find work.’			
Not Actionable	-74.1179, 18.6423	1/22/2010 18:29	5. Other
Rue Casseus no 9 gen yon sant kap bay swen ak moun ki blese e moun ki brile.			
‘Street Casseus no 9, there is a center that helps people that are wounded or burnt.’			
Actionable	-72.32857,18.53019	1/19/2010 11:21	4a. Health services
Paket moun delmas 32 ki forme organisation kap mache pran manje sou non pep yo ankesel lakay			
‘People in Delmas 32 formed an association that takes food in the name of everyone in the neighborhood’			
Not Actionable	-72.30815,18.54414	2/4/2010 1:21	5. Other

Table 1: Examples of messages, with translation, actionable flag, location, timestamp and categories. The final two were a consistent false negative and false positive respectively. The former likely because of sparsity - places offering services were rare - and the latter because reports of possible corruption were not, unfortunately, considered actionable.

A set S, of 49 base streaming models predicted the has-location, $L_{\exists}$, feature and categories, $c_{0,\ldots,47}$. That is, unlike the main model, R, which predicts the ‘Actionable’/‘Not Actionable’ division, each model S predicts either the existence of a location within the text or binary per-category membership. The predictions for each message were added to the final model as feature confidence values per-label. This is richer than simply using the best-guess labels for each model S and it was clear in early processing that confidence features produced consistently more accurate final models than binary predicted labels. In fact, using model confidence features for $L_{\exists}$ actually outperforms an oracle binary has-location feature, $O(L_{\exists})$ (see results).

The same G, W, P and T features were used for the S and R models.

As the final model R requires the outputs from S for training, and S are themselves predictive models, the $L_{\exists}$ and C features are not included until the second training epoch R_1. In the context of machine-learning for sudden onset humanitarian information processing any delay could be significant. One simple solution is starting with smaller epoch sizes.

4.2 Subword features

A typical former creole, Haitian Kreyol has very simple morphology but the text message-language produces many compounds and reductions (‘my family’: *fanmi mwen, fanmwen, fanmi m, fanmi’m, fanmim’, fanmim*), so it requires segmentation. There is also substantial variation due to lack of spelling conventions, geographic variation, varying literacy, more-or-less French spellings for cognates, and character sets/accents (‘thank you’: *mesi, mési, méci meci, merci*). See Table 2 for further examples and common subword patterns that were discovered across very different surface forms.

The approach here builds on the earlier work of Munro and Manning (2010), adapted from Goldwater et al. (2009), where unsupervised methods were used to segment and phonologically normalize the words. For example, the process might turn all variants of ‘thank you’ into *‘mesi’* and all variants of ‘my family’ into *‘fan mwen’*. This regularization allows a model to generalize over the different

Abbrev.	Full Form	Pattern	Meaning
s'on	se yon	*sVn*	is a
avèn	avèknou	*VvVn*	with us
relem	rele mwen	*relem*	call me
wap	ouap	*uVp*	you are
map	mwen ap	*map*	I will be
zanmim	zanmi mwen	*zanmim*	my friend
lavel	lave li	*lavel*	to wash (it)

Table 2: Abbreviations and full forms of words, showing substantial variation but common subword patterns and character alternations (V=any vowel).

forms even in the event of singleton word variants. Here we incorporated the segmentation into the supervised learning task rather than model the phonological/orthographic variation as a pre-learning normalization step, as in Munro and Manning. A set of candidate segmentations and normalizations were added to our final model as features representing both the pre and post-normalization, allowing the model to arrive at the optimal training weights between the unnormalized/unsegmented and normalized/segmented variants.

This meant that rather than optimizing the subword segmentation according to a Gaussian prior over unlabeled data we optimized the segmentation according to the predictive ability of a given segmentation *per model*. This is further from the linguistic reality of the segments than our earlier approach but the richer feature space led to an increase in accuracy in all test cases here.

The subword models were only applied to the original messages. The English translations were not included among the features as it is not realistic to assume manual translation of every message before the less labor-intensive task of identifying actionable items.

4.3 Oracle features

While the SMS protocol does not encode the geographical origin of the message, other protocols/systems do, especially those used by smartphones. Similarly, cell-tower granularity of locations might be available, phone numbers might be *a priori* associated with locations, or locations could be formalized within the text using methods like 'Tweak the Tweet' (Starbird and Stamberger, 2010). Therefore, it is reasonable to also simulate a scenario where the messages come pre-geocoded. We compared our results to models also containing the oracle longitude and latitude of the messages, $O(L)$ (no attempt was made to predict L, the precise longitude and latitude - a challenging but interesting task) and the oracle existence of a location $O(L_{\exists})$.

It is less likely that messages come pre-categorized but oracle features for the categories were also evaluated to compare the performance for models containing the predictive categories to ones containing the actual categories, $O(c_{0,\ldots,47})$.

4.4 Spatial clustering

In addition to identifying locations, we also used the latitude and longitude to geographically cluster messages. This was to capture two phenomena:

1. *Hot spots:* some areas were in greater need of aid than others.

2. *Clustered food requests:* the definition of 'actionable' extended to clustered requests for food, but not requests from lone individuals.

Figure 1 shows a Port-au-Prince (Pótoprens) neighborhood with incident reports from the text messages. The x, y axes (latitude, longitude) show the clusters given by the Ushahidi map and the z axis shows the temporal distribution of messages over a two month period. Both the spatial and temporal distributions clearly show a high frequency of both clusters and outliers.

The most accurate clustering divided the messages by longitude and latitude into tiles approximating $100m^2$, $1km^2$ and $10km^2$. At each granularity, tiling was repeated with an offset by half on each dimension to partially smooth the arbitrariness of tile boundaries. This resulted in each geolocated messages being a member of 12 tiles in total, which were included as 12 features L. We were not able to find an unsupervised spatial clustering algorithm that improved the results beyond this brute-force method of multiple tiles at different granularities (see discussion of other models tested below).

4.5 Temporal modeling and discounting

It is common to calculate a discounting function over training epochs in streaming models (Aggar-

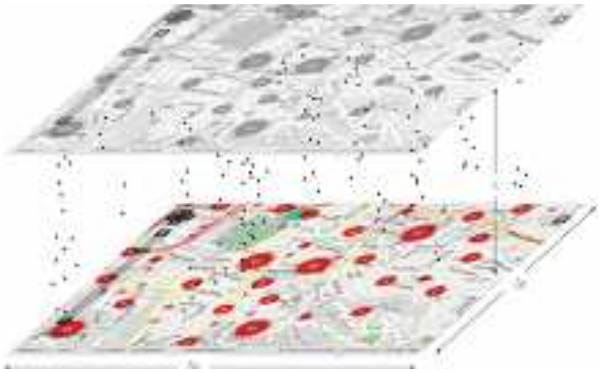

Figure 1: Map of a Port-au-Prince neighborhood with incident reports from text messages, with spatial clusters on the latitudinal and longitudinal axes and temporal distributions on the time axis, showing both spatial and temporal clustering, with frequent outliers.

wal, 2006; Hulten et al., 2001; Babcock et al., 2002).

We used a slightly different method here where the time-stamp feature, T, performs this function, arriving at the relative probability for a given time period t in the final model, R (Zhang et al., 2008). It has several advantages over a simple weighted discounting function. First, t is calculated incrementally, not per training epoch, meaning that the weight θ for t is calculated until the most recently seen items. Second, it frees T to cluster according to temporal divisions other than the (arbitrary) divisions of training epochs. Finally, it allows unconstrained weights for different temporal clusters, permitting the final distribution of weights over different ts to define complex and possibly nonmonotonic discounting functions. Modeling time as a feature rather than a discounting function also made it simpler to combine temporal and spatial clustering. The feature T consisted of multiple buckets of time-stamps per message and also composite features with the $O(L)$ tiles when present.

4.6 Other models tested

Several other machine-learning methods were tested but ultimately not reported here.

Intuitively, SVMs with non-linear kernels could more accurately model geographic divisions in the latitude and longitude dimensions and discover different combinations of features like *has-location=true* and *category=emergency*. However, we were not able to find a combination of kernels and parameter settings that demonstrated this. It is possible that we could not avoid over-fitting or that the composite features had already sufficiently captured the combinations.

Munro and Manning (2010) also found gains using supervised LDA (Ramage et al., 2009), which has also previously been used for disaster response clustering (Kireyev et al., 2009). We implemented supervised LDA and unsupervised LDA topic models, but they showed modest improvements over the baseline model only. We presume that this is because when we add the predicted categories from our (supervised) category learning task, they already contained enough information about topic divisions.

We looked at several methods for spatio-temporal clustering including cliques (Zhang et al., 2004), k-means (Wagstaff et al., 2001), and targeted low frequency clusters (Huang et al., 2003). The change in accuracy was negligible but the exploration of methods was by no means exhaustive. One exception was using nearest-neighbor spatiotemporal clustering, however the gains were predominantly repeated messages from the same person and thus already captured by the source feature, P.

Several systems have been built by humanitarian organizations for filtering/prioritizing messages, mostly keyword and memory-based. All are currently less accurate than the baseline system here and their predicted outputs gave no gains as features. The *SwiftRiver* system built on the NLTK library (Bird et al., 2009) was the most promising, only underperforming the baseline by F=0.01.

5 Results

The results in Table 3 show that a combination of streaming models with subword models improves the accuracy of identifying actionable messages. All increases in accuracy are significant relative to the baseline.

The temporal feature, T, alone gives F=0.045 increase in accuracy indicating that there is substantial concept drift and an adaptive model is necessary for accurate classification.

The subword models, W, increase the gain to 0.119 showing that despite Kreyol being a morphologically simple language the variation in spellings and compounds is significant.

Model	Precision	Recall	F-value	F-Gain
Words/Ngrams (G)	0.622	0.124	0.207	*n/a*
Temporal feature (G, T)	0.716	0.153	0.252	0.045
Subwords and Source (G, T, W, P)	0.548	0.233	0.326	0.119
Categories (predicted: G, T, C, P)	0.464	0.240	0.316	0.109
Location (predicted: $G, T, L_{\exists}, P$)	0.572	0.212	0.310	0.103
Categories (oracle: $G, T, O(C), P$)	0.565	0.225	0.322	0.115
Location (oracle: $G, T, O(L_{\exists}), P$)	0.746	0.168	0.274	0.067
Spatial clusters (L)	0.896	0.653	0.756	0.549
All non-oracle and spatial clusters	0.872	0.840	**0.855**	0.648
Pre-filtered spatial clusters	0.613	0.328	0.428	0.221
Radio station	0.961	0.854	0.904	*n/a*
Twitter	0.950	0.989	0.969	*n/a*

Table 3: The final results for the different models. The first three and *Location (oracle)* contain only single streaming models. The others use a hierarchical streaming model combining features with the outputs from the base streaming models S. The model combining all features is the most accurate at F=0.855, 0.648 above the baseline model optimized over words and word sequences (ngrams). The *Pre-filtered spatial clusters* contains the same architecture/features as the most accurate model, but with those messages not containing an identifiable location (and therefore automatically non-actionable) stripped from the training and test data. The final *Radio station* and *Twitter* models used the messages sent to a radio station and Twitter as the non-actionable items, using the same training model as *All non-oracle and spatial clusters*.

5.1 Oracle vs predicted outputs

Comparing the oracle values and predicted outputs from the categories and the identification of messages containing locations, $O(C), O(L_{\exists}), C, L_{\exists}$, we see that the predictive model for categories only slightly underperforms the oracle, but the predictive model for locations *out*performs the oracle model. We suspect that this is because the predictions are probabilities, not binary indicators of the existence of a location. Therefore, the richer real-valued feature space led to greater information for the final model despite any predictive errors in this base model. Another reason can be clearly seen in the precision, 0.746 for the $O(L)$ model, one of the highest precision-to-recall ratios. The final model is clearly giving too much weight to the existence of a location as a predictor of an actionable label. Smoothing $O(L)$ makes little difference as it is a high-frequency binary feature, but the greater range of probability values in L are necessarily more sparse and therefore more de-weighted by smoothing.

Identifying locations is one area that could be expanded on greatly. Initial experiments with named-entity recognition were abandoned when it was clear that the data was too sparse and too different from existing data sets, but perhaps text-message-specific named-entity recognition methods could lead to even more accurate results for identifying locations.

5.2 Spatial clustering

By adding spatial clusters we get the greatest leap in accuracy: F=0.756, a substantial F=0.549 over the baseline. This is strong evidence in favor of extending text-only messaging to location-aware messaging. As with the prediction of locations, it is likely that methods for spatial clustering other than brute-force bucketing could lead to more accurate results, but as stated earlier we were not able to identify any.

Combining the base streaming model outputs with all features leads to the most accurate model. It is expected that this would produce the best results, but at F=0.855 we have a *very* significant gain over any of the models implementing only single-stream learning, or without the full feature space.

5.3 Pre-filtering

Somewhat counterintuitively, pre-filtering messages without known locations (in both training and test data) decreased the accuracy to F=0.428. Oracle filtering of true-negative test items will not change recall and can only increase precision, so clearly there is 'signal' in the 'noise' here. Analysis of the messages showed that many non-actionable messages were not related to emergencies at all (general questions, requests for work, etc), as were many messages without identifiable locations. That is, people who tended to not send actionable information also tended to not include locations. Because of this correlation, the content of messages without locations becomes useful information for the model.

A careful analysis of the training models confirms this: the word and subword features for non-actionable messages with no locations had non-zero weights. Pre-filtering them therefore resulted in an impoverished training set.

It is standard practice in the humanitarian industry to pre-filter messages that are easily identified as non-actionable (for obvious reasons: it reduces the manual processing), which is what occurred in Haiti - only about 5% of messages were treated as 'actionable' candidates. The results here indicate that if manual processing is extended to automated or semi-automated processing this strategy needs to change, with all potential training items included in the models.

5.4 Comparison to non-emergency messages

For the social media scenarios where we combined the actionable test items with the messages sent to a radio station and Twitter the accuracy was highest of all. This is a promising result for seeking actionable information in non-traditional sources. The radio station is particularly promising as almost all the messages where in Haitian Kreyol and spoke about the same locations as the 4636 messages.

While the Twitter messages were extremely accurate at F=0.969, the majority of the tweets were in English or French from people outside of Haiti, so this model was at least in part about language identification, a much simpler task and less novel from a research perspective. Nonetheless, while at least part of the accuracy is easily explained this was the most sparse test set with only 0.025% actionable items, so the application scenario is very promising.

5.5 Prioritization

Applying ROC analysis, the methods here could speed up the prioritization of actionable messages by a factor of 10 to 1 based on content alone. That is, on average an actionable message falls within the 90th percentile for probability of being actionable. By including spatial clustering this becomes the 98th percentile. There is great potential for improvements but the methods reported here could already be used to efficiently prioritize the triage of the most important messages within a semi-automated system.

6 Related Work

6.1 trillion text messages were sent in 2010 - more than emails and social network communications combined (ITU, 2010), especially in areas of great linguistic diversity (Maffi, 2005). This easily makes it the *least* well-studied method for digital communication relative to the amount digital information being generated.

The lack of research is probably due to obstacles in obtaining data. By contrast Twitter's API has led to much recent research, primarily in sentiment analysis (O'Connor et al., 2010; Alexander Pak, 2010; Sriram et al., 2010) and unsupervised event detection (Petrović et al., 2010). The task of identifying sentiment is different to filtering actionable intelligence, we were not training on tweets, and Twitter-language is reportedly different from text-message-language (Krishnamurthy et al., 2008). However, there are similarities relating to problems of message brevity and the ability to extend the feature-space. For example, Sriram et al. (2010) also found that modeling the source of a message improved accuracy. Eisenstein et al. (Eisenstein et al., 2010) show promising results in identifying an author's geographic location from micro-blogs, but the locations are course-grained and rely on a substantial message history per-source.

In recent work with medical text messages in the Chichewa language, we compared the accuracy of rule-based and unsupervised phonological normalization and morphological segmentation when applied to a classification task over medical labels,

showing substantial gains from subword models (Munro and Manning, 2010).

A cluster of earlier work looked at SMS-SPAM in English (Healy et al., 2005; Hidalgo et al., 2006; Cormack et al., 2007) and Beaufort et al. (2010) used a similar preprocessing method for normalizing text-messages in French, combining rule-based models with a finite-state framework. The accuracy was calculated relative to BLEU scores for 'correct' French, not as a classification task.

Machine-translation into a more well-spoken language can extend the potential workforce. Early results are promising (Lewis, 2010) but still leave some latency in deployment.

For streaming architectures, Zhang et al. (2008) proposed a similar method for calculating per epoch weights as an alternative to a discounting function with significant gains. Wang et al. (2007) also looked at multiple parallel streams of text from different newspapers reporting the same events but we couldn't apply their method here as there were few instances of the same pairs of people independently reporting two distinct events. The two-tiered architecture used here is similar to a hierarchical model, the main difference being epoch-based retraining and the temporal offset of the base models feeding into the final one. Joint learning over hierarchical models has been successful in NLP (Finkel and Manning, 2010) but to our best knowledge no one has published work on joint learning over hierarchical streaming models, in NLP or otherwise.

7 Conclusions

From models optimized over words and ngrams to one including temporal clustering and subword models the accuracy rises from F=0.207 to F=0.326. Clearly, the words that someone has chosen to express a message is just one small aspect of the context in which that message is understood and by combining different learning models with richer features we can prioritize actionable reports with some accuracy. With spatial clustering this rises to F=0.885, indicating that geographic location was the single most important factor for prioritizing actionable messages.

These results are only a first step as there is great potential for research identifying more accurate and efficient learning paradigms. A growing number of our communications are real-time text with frequent spelling variants and a spatial component (tweets, location-based 'check-ins', instant messaging, etc) so there will be increasingly more data available in an increasing variety of languages.

It is easy to imagine many humanitarian applications for classifying text-messages with spatiotemporal information. Social development organizations are already using text messaging to support health (Leach-Lemens, 2009), banking (Peevers et al., 2008), access to market information (Jagun et al., 2008), literacy (Isbrandt, 2009), and there is the potential to aid many of them. Even more importantly, this work can contribute to information processing strategies in future crises. Had a system like the one presented here been in place for Haiti then the identification of actionable messages could have been expedited considerably and a greater volume processed. I coordinated the Mission 4636 volunteers who were translating and mapping the messages in real-time, so this research is partially motivated by the need to see what I could have done better, with a view to being better prepared for future crises.

The results for social media are especially promising. In total, the tweets contained 1,178,444 words - the size of approximately 10 novels - but if there was just one real emergency among them, there was a 97% chance it would rise to the top when ordered by actionable confidence.

Acknowledgments

With thanks to the volunteers of Mission 4636. Their work translating, categorizing and mapping communications showed the humanitarian community the benefits of crowdsourcing/microtasking and is now also helping us prepare for higher-volume semi-automated systems. Thanks also to the volunteers and workers of Samasource/FATEM in Haiti, Ushahidi Haiti in Boston, and to the engineers at CrowdFlower and Ushahidi who built the platforms we used for translation and mapping.

This work was supported by a Stanford Graduate Fellowship and owes thanks to collaborative work and conversations with Chris Manning and colleagues in the Stanford NLP Research Group.

References

Charu C. Aggarwal. 2006. *Data Streams: Models and Algorithms (Advances in Database Systems)*. Springer-Verlag, New York.

Patrick Paroubek Alexander Pak. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceeding of the 2010 International Conference on Language Resources and Evaluation (LREC 2010)*.

Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. 2002. Models and issues in data stream systems. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–16. ACM.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL 2010*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly & Associates Inc.

David A. Cohn, Zoubin Ghahramani, and Michael Jordan. 1996. Active learning with statistical models. *Arxiv preprint cs/9603104*.

Gordon V. Cormack, José Mara Gómez Hidalgo, and Enrique Puertas Sánz. 2007. Feature engineering for mobile (SMS) spam filtering. In *The 30th annual international ACM SIGIR conference on research and development in information retrieval*.

The American Red Cross. 2010. Social media in disasters and emergencies. Presentation.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*.

Jenny Rose Finkel and Christopher D. Manning. 2010. Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *Annual Conference of the Association for Computational Linguistics (ACL 2010)*.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Matt Healy, Sarah Jane Delany, and Anton Zamolotskikh. 2005. An assessment of case-based reasoning for Short Text Message Classification. In *The 16th Irish Conference on Artificial Intelligence & Cognitive Science*.

José Mara Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, and Francisco Carrero Garca. 2006. Content based SMS spam filtering. In *ACM symposium on Document engineering*.

Yan Huang, Hui Xiong, Shashi Shekhar, and Jian Pei. 2003. Mining confident co-location rules without a support threshold. In *Proceedings of the 2003 ACM symposium on Applied computing*.

Geoff Hulten, Laurie Spencer, and Pedro Domingos. 2001. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Scott Isbrandt. 2009. Cell Phones in West Africa: improving literacy and agricultural market information systems in Niger. White paper: Projet Alphabétisation de Base par Cellulaire.

ITU. 2010. The world in 2010: ICT facts and figures. *International Telecommunication Union*.

Abi Jagun, Richard Heeks, and Jason Whalley. 2008. The impact of mobile telephony on developing country micro-enterprise: A Nigerian case study. *Information Technologies and International Development*, 4.

Kirill Kireyev, Leysia Palen, and Kenneth M. Anderson. 2009. Applications of topics models to analysis of disaster-related Twitter data. In *Proceedings of the NIPS Workshop on Applications for Topic Models: Text and Beyond*.

Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. 2008. A few chirps about Twitter. In *Proceedings of the first workshop on Online social networks*, New York.

Carole Leach-Lemens. 2009. Using mobile phones in HIV care and prevention. *HIV and AIDS Treatment in Practice*, 137.

Will Lewis. 2010. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In *14th Annual Conference of the European Association for Machine Translation*.

Luisa Maffi. 2005. Linguistic, cultural, and biological diversity. *Annual Review of Anthropology*, 34:599–617.

Robert Munro and Christopher D. Manning. 2010. Subword variation in text message classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*.

Robert Munro. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010.

From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Gareth Peevers, Gary Douglas, and Mervyn A. Jack. 2008. A usability comparison of three alternative message formats for an SMS banking service. *International Journal of Human-Computer Studies*, 66.

Sasa Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore.

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

Kate Starbird and Jeannie Stamberger. 2010. Tweak the Tweet: Leveraging Microblogging Proliferation with a Prescriptive Syntax to Support Citizen Reporting. In *Proceedings of the 7th International ISCRAM Conference*.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. 2001. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, volume 577, page 584.

Xuanhui Wang, Cheng Xiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Xin Zhang, Nikos Mamoulis, David W. Cheung, and Yutao Shou. 2004. Fast mining of spatial collocations. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 384–393. ACM.

Peng Zhang, Xingquan Zhu, and Yong Shi. 2008. Categorizing and mining concept drifting data streams. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.