Amplifying the Voice of Youth in Africa via Text Analytics

Prem Melville
IBM Research
pmelvil@us.ibm.com

Vijil Chenthamarakshan
IBM Research
ecvijil@us.ibm.com

Richard D. Lawrence
IBM Research
ricklawr@us.ibm.com

James Powell
UNICEF Uganda
james123powell@hotmail.com

Moses Mugisha
UNICEF Uganda
mossplix@gmail.com

ABSTRACT

U-report is an open-source SMS platform operated by UNICEF Uganda, designed to give community members a voice on issues that impact them. Data received by the system are either SMS responses to a poll conducted by UNICEF, or unsolicited reports of a problem occurring within the community. There are currently 200,000 U-report participants, and they send up to 10,000 unsolicited text messages a week. The objective of the program in Uganda is to understand the data in real-time, and have issues addressed by the appropriate department in UNICEF in a timely manner. Given the high volume and velocity of the data streams, manual inspection of all messages is no longer sustainable. This paper describes an automated message-understanding and routing system deployed by IBM at UNICEF. We employ recent advances in data mining to get the most out of labeled training data, while incorporating domain knowledge from experts. We discuss the trade-offs, design choices and challenges in applying such techniques in a real-world deployment.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining—*Machine Learning*; H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Machine Learning, Text Classification, Information Retrieval

1. INTRODUCTION

In May 2011, UNICEF Uganda launched the U-report program to allow young people in Uganda to voice opinions on a range of social issues that impact their daily lives. Young people participate by first enrolling in the program, and then texting (SMS) responses to frequent polls created by UNICEF. Recent polls have asked questions such as "Have you heard any immunisation adverts on the radio?" and "What's the most important problem u want the government to solve to make life better for you and your family?". Poll responses are automatically parsed and summarized by geographical location on the U-report website (http://www.ureport.ug). Currently, over 200,000 young people in Uganda have enrolled in U-report, and the program is adding 200 to 1000 new participants every day. U-report is a novel program in that it allows young people living below the poverty line in developing countries to communicate with governmental organizations using the dominant means of communication (SMS) readily available to them.

Participants (known as U-reporters) can also text unsolicited reports of a problem in their community. These messages include concerns and observations about health care, education, gender violence and other issues, some of which may require immediate action from UNICEF or the local government. Approximately 10,000 such messages are received each week. Of these messages, approximately 39% require an SMS response providing advice or an answer to a question (classified as *informational*) and 7% required an immediate action or intervention (classified as *actionable*) by either a government stakeholder, a non-government organization (NGO), or by UNICEF themselves. Given the sheer volume of received texts, coupled with the potential urgency of the messages, UNICEF Uganda joined forces with IBM Research to develop an automated text classification system to facilate routing of the received messages to the responsible agency.

This paper describes such a system, which has been deployed within the open-source SMS platform operated by UNICEF Uganda. We describe the overall system in the next section, and then discuss data pre-processing in Section 3. Sections 4 through 6 describe the specific text analytics and machine learning approaches, Section 7 discusses the deployment, with concluding remarks in Section 8.

2. SYSTEM OVERVIEW

The U-report application and database are hosted in Uganda on United Nations servers. (UNICEF is an agency within the United Nations.) Figure 1 shows a high-level view of the system. U-reporters send SMS messages that are collected via a third-party SMS aggregator and sent to the U-report SMS gateway. The texts, along with meta-data and profile information, are stored in the U-report database. These texts can include responses to polls as well as unsolicited messages, and hence the first text analytics task is to identify and parse poll responses. This filtering step identifies poll responses based on the presence of anticipated response patterns (e.g. `yes` or `no`), the length of the message (short

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

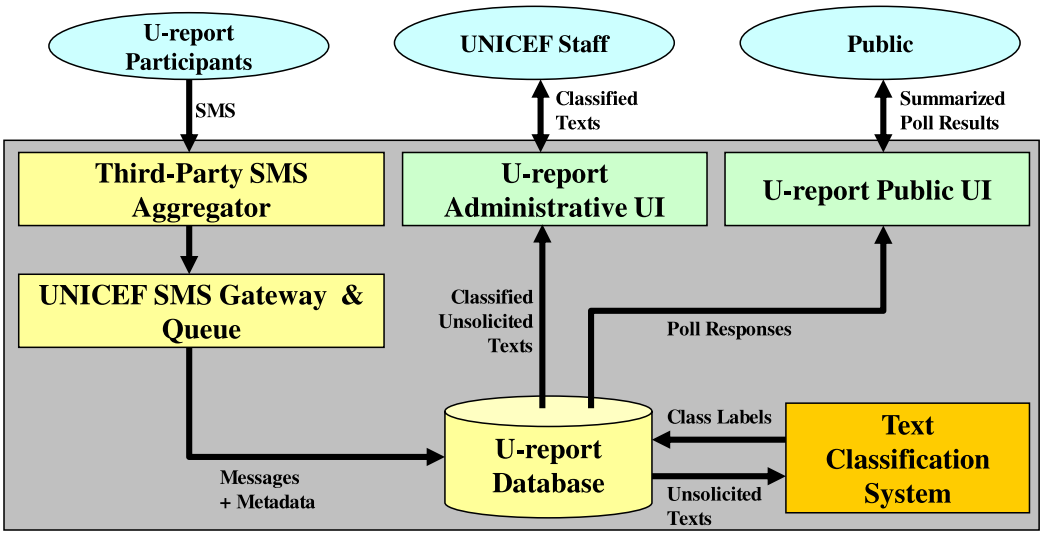

Figure 1: Overview of U-report System

messages are more likely to be poll responses), and the timestamp of the message to match it to a recent poll question. The parsed poll responses are stored in the database, and summarized by geography via the public UI on the U-report website. In this paper, we focus on the methodology developed to classify the unsolicited texts, after the poll responses have been filtered as described above. The classified unsolicited texts are made available to UNICEF staff via a limited-access administrative interface (discussed further in Section 7). Based on this classification, ranked lists of messages can be routed to specific UNICEF teams focusing on Education, Health, or Child Protection.

3. DATA AND METHODOLOGY

In this section we describe the UNICEF datasets, preprocessing steps, and our evaluation methodology.

UNICEF Uganda volunteers classified a sample of historical text messages into 1 of the 13 categories described below:

- ***water***: Includes messages about water, hygiene and sanitation.
- ***health & nutrition***: Includes messages about HIV, AIDS, breast feeding, malaria, other illnesses and food shortages.
- ***orphans & vulnerable children***: Includes messages with references to child labor, orphans, early marriage, domestic violence and teenage pregnancy.
- ***violence against children***: Messages about violence targeting children in schools and at home.
- ***education***: Includes messages with references to schools or any learning environment for children and young people.
- ***employment***: Includes messages about inflation, unemployment and youth finance issues.
- ***emergency***: Message regarding pressing matters such as disease outbreaks, landslides, and refugee situations.
- ***social policy***: Messages relating to critical policy issues, such as the Children Act.
- ***u-report***: Messages praising or criticizing the U-report program.
- ***energy***: Messages about energy challenges, shortages and innovations, e.g., biogas, solar, etc.
- ***family & relationships***: Messages about issues at home, normally involving relationships.
- ***irrelevant***: Messages unrelated to development.
- ***poll***: Responses to polls (typically `yes` or `no`).

The labeled historical data allows us to evaluate our routing system, as well as provides training data for our supervised models. While each message was labeled with 1 of 13 labels, in practice messages could be relevant to more than one class. So, rather than treat this as a 13-way classification problem, we view it as multiple binary-classification tasks; where each classifier determines if a message is relevant to a particular class or not. We do not explicitly model the *irrelevant* category, as a message that has a low relevance score on all classes is automatically deemed as irrelevant. We also do not explicitly model the *poll* class, since messages that are poll responses are handled outside of the Text Classification System as noted in the previous section. This leaves us with 11 binary classification problem, for which we have a total of 54,561 training examples. These binary-class datasets are highly skewed, since only between 0.5 and 11% of the messages are relevant to a particular class; and 68% of the examples are not relevant to any of the classes. For the supervised text classifiers we also convert this data into word-vectors, represented by term frequencies in each message, after eliminating stop-words.

The output of our system are lists of messages ranked by relevance to each of the 11 categories. Since we are primarily concerned with ranking, we evaluate different approaches based on area under ROC curves (AUC). All our experimental results are based on 10-fold cross-validation, and statistical significance is determined by paired t-test ($p < 0.05$).

Note that, the official language of Uganda is English and around 95% of the messages are in English. A U-reporter enrolls in the system by sending a specific keyword (`join`, `donyo`, or `togeu`) in one of the three popular languages (English, Luo, and Karamojong repectively). The specific keyword used by the U-reporter is used to identify his primary language and is stored in his profile. We currently focus only on the messages sent by U-reporters whose primary language is English. Providing multi-lingual support is a challenge to be tackled in future deployements.

4. KEYWORD MATCHING VERSUS CLASSIFICATION

With the rapid adoption of the U-report program, manual inspection of unsolicited message are no longer feasible. In this section we discuss some initial attempts at automating the process of detecting relevant messages, and some improvements on these baselines. The results of the approaches we describe below are all summarized in Table 1.

4.1 Keyword Matching

In the process of manually inspecting messages, UNICEF employees found several commonly occurring terms in different categories. An obvious first attempt at automation was to create lists of terms relevant to each class, and to match messages against this list.

UNICEF created lists of n-grams indicative of each of the 11 classes. For instance, the list for *health & nutrition* includes medical terms such as, `AIDS, HIV, malaria, polio, TB, doctor, nurse, drugs, hospital,` etc. In addition, the list also contains context-specific terms such as `HC 1, HC 2, ... HC 5` (referring to different levels of healthcare centers in Uganda) and `VHT` (referring to the `voluntary health team`).

In order to evaluate this keyword matching approach, we consider every message with the presence of at least one keyword to be relevant to the class under consideration, and the rest to be irrelevant. The subsequent labeling on test data can be evaluated using area under ROC curves, similar to supervised classifiers. The results of this approach can be seen in the second column of Table 1. This simple approach does work reasonably well for some categories, such as *water* and *energy*. However, other categories are not as easy to identify with predetermined keyword lists alone. For instance, the list for *social policy*, which includes `policy, government, election, corruption`, etc., is clearly insufficient to capture the broad concept of messages calling for reform or discussing critical policy issues in the region.

4.2 Text Classification

Given the inadequacy of using predetermined keyword lists, an alternative is to train a text classifier based on labeled data. In particular, we apply Naïve Bayes and SVMs to the word-vector data described in Section 3. Here, supervised learning lets us detect patterns in our data that the human experts were unable to identify. For instance, here are the top 20 terms that are most discriminative of the *violence against children* class in the data, as determined by χ^2 scores [7]: `defilement, child, female, fgm, sacrifice, defiled, raped, abuse, circumcision, girl, violence, genital, rape, mutilation, practice, defile, cases, defiling, man, beaten`. Only 7 of these 20 words were identified by experts in their list of relevant keywords.

The text classification results are summarized in columns 3 and 4 of Table 1. While, on average, the performance of the text classifiers is better, there are several datasets for which keyword matching is still statistically equivalent or better.

4.3 Domain Challenges

While, the initial results of applying text classification may not seem very promising, a closer examination reveals that the problem lies in the quality of data. Since the text in this domain comes from SMS messages, and English proficiency is not always high amongst the user base, the messages often contain various abbreviations, and are rife with spelling errors. In order to deal with this, we implemented a process to auto-correct words that are greater than 3 letters in length. The process leverages a combination of Philips' metaphone algorithm and string-edit distance.[1] For spell-correction, in addition to a general-purpose English dictionary, we also use a dictionary of terms specific to Uganda, such as local slang and the list of all districts in the country. Furthermore, we normalize abbreviations common in texting to their canonical form, so that, e.g., `b4` and `be4` are both mapped to `before`.

Note that auto-correction involves automatically picking the best match for each word, since there is not a human in the loop to select from the possible alternatives. As such, the process is far from perfect, but it does allow us to identify relevant patterns that may have otherwise gone unnoticed. Consider, for example, the messages below, before (B) and after (A) auto-correction:

B: `hallo this is <anon> in community people are suffuring from maleria`
A: `hallo this is <anon> in community people are suffering from malaria`

B: `by the way; is there a way we stop maleria in pece? if so then why is it not happing`
A: `by the way is there a way we stop malaria in Peace if so then why is it not harping`

B: `people are facing the problem maleria esp childen (ntungamo)what can we do?`
A: `people are facing the problem malaria esp child en noncom what can we do`

While some auto-corrections are clearly erroneous (`harping` for `happing`), we are able to correctly classify the above messages as relevant to *health & nutrition* based on the corrected mentions of malaria.

After this auto-correction process, we re-ran the Naïve Bayes and SVM classifiers, and report the results in column 6 and 7 in Table 1. We now see a substantial improvement over the baseline, where Naïve Bayes is statistically significantly better on 8 of the 11 datasets.

Note that spell-correction also helps keyword-based classification on 5 datasets, but not on average. This is because the UNICEF keyword lists already contain several common misspellings of relevant terms. So the improvements we see in the text classifiers are from identifying novel patterns, and correcting for more egregious spelling errors than the keyword matching process. After correcting for errors, Naïve Bayes tends to outperform SVMs on this data, so we use it as our supervised baseline for the rest of the paper.

[1] http://aspell.net/

	Original Messages			Spell-Corrected		
Dataset	Keywords	Naïve Bayes	SVM	Keywords	Naïve Bayes	SVM
education	79.8	78.7	77.2	80.1	**86.9**	77.6
emergency	60.8	60.8	**67.8**	61.6	**69.7**	**69.1**
employment	80.5	67.0	78.9	**81.3**	77.8	79.8
energy	90.0	78.3	88.2	72.2	85.2	89.4
family	55.3	54.3	**66.2**	55.9	**64.9**	**67.5**
health	77.5	**84.5**	**81.0**	77.6	**89.8**	**82.2**
orphans	56.4	**65.3**	**68.2**	**57.7**	**75.9**	**69.0**
social	58.8	60.7	**66.3**	**62.6**	**71.8**	**66.8**
u-report	64.8	**88.2**	**76.8**	**65.1**	**90.0**	**76.5**
violence	64.7	65.2	68.5	**65.9**	**75.4**	**70.5**
water	94.4	89.0	91.8	94.4	94.0	92.3
Mean	71.2	72.0	75.5	70.4	80.1	76.4
Sig. Win/Draw/Loss		3–5–3	6–3–2	5–5–1	8–2–1	7–2–2

Table 1: Performance in terms of AUC of baseline keyword matching approach, text classifiers and spell-correction. Statistically significant improvements over Keywords (Column 2) are presented in bold.

5. DUAL SUPERVISION

Keyword lists, as used in Section 4.1 can be viewed as one form of background knowledge about the domain. Such domain knowledge need not be discarded in lieu of using a text classifier. We recently introduced the *dual supervision* framework [5, 6], in which classifiers are trained using both labels on instances, as well as such prior knowledge on associations of features (words) to particular classes. We refer to word-class associations as *word labels* or *feature labels*, drawing a parallel to document or instance labels. In this section, we explore the use of an approach to learning from both feature labels and instance labels, in an attempt to further improve our classifications.

5.1 Pooling Multinomials

The Pooling Multinomials classifier [5] was introduced as an approach to incorporate prior lexical knowledge into supervised text classification for improved sentiment analysis. In the context of sentiment analysis, such lexical knowledge is readily available in terms of the prior sentiment-polarity of words. However, the same approach is applicable to any text classification setting, where prior knowledge is available about the class-association of some terms in a domain. In our current domain, the keyword lists serve as a mapping between some words and the class they are indicative of, and hence can be treated as feature labels, e.g., `borewell` has the label *relevant* for the *water* dataset.

Pooling Multinomials classifies unlabeled examples just as in multinomial Naïve Bayes classification, by predicting the class with the maximum likelihood, given by

$$argmax_{c_j} P(c_j) \prod_i P(w_i|c_j)$$

where $P(c_j)$ is the prior probability of class c_j, and $P(w_i|c_j)$ is the probability of word w_i appearing in a document of class c_j. In the absence of background knowledge about the class distribution, we estimate the class priors $P(c_j)$ solely from the training data. However, unlike regular Naïve Bayes, the conditional probabilities $P(w_i|c_j)$ are computed using both labeled examples and labeled features. Given two models built using labeled examples and labeled features, the multinomial parameters of such models can be aggregated through a convex combination,

$$P(w_i|c_j) = \alpha_1 P_e(w_i|c_j) + \alpha_2 P_f(w_i|c_j)$$

where $P_e(w_i|c_j)$ and $P_f(w_i|c_j)$ represent the probability assigned by using the example labels and feature labels respectively, and α's are weights for combining these distributions. The derivation and details of these models are not directly relevant to this paper, but can be found in [5].

The weights in the equation above, indicate a level of confidence in each source of information, and Melville et al. [5] set these automatically based on the training set accuracy of each component. Since we are primarily concerned with ranking in our domain, we modified Pooling Multinomials to select weights based on AUCs on the training set. In particular we use a sigmoid weighting scheme:

$$\alpha_k = \log \frac{auc_k}{1 - auc_k}$$

where auc_k is the area under the ROC curve of model k on the training set; and the α_k's are normalized to sum to one. We refer to this variant of Pooling Multinomials as *Pooling-AUC*. We also experimented with explicitly setting the weights to be equal, making the simplifying assumption that instance and feature labels are equally valuable. We will refer to this default weight setting simply as Pooling Multinomials.

5.2 Design Choices

The Pooling Multinomials algorithm expects labeled instances as well as labeled features for both classes. As before, we use instances from the target class as the *relevant* instances, and instances from all other classes as the *irrelevant* instances. Analogously, we use the keyword list for one class to generate the feature labels for the *relevant* class, and the keyword lists of all other classes to generate the feature labels for the *irrelevant* class. Note that keywords for other classes are not necessarily indicative of irrelevance to a given class. Also, a message can belong to more than one class. However, in practice, this approach of creating feature labels works quite well. Given enough labeled messages, Pooling Multinomials is able to correct for discrepancies in the *irrelevant* feature labels.

Table 2 summarizes the results of dual supervision versus using only labeled instances. We see that by incorporating the background domain knowledge through Pooling Multinomials we are able to achieve higher AUCs on 8 of the datasets (statistically significant in 6 cases). We typically see a 1 to 5% increase in AUC, corresponding to significant improvements in identifying relevant messages. We observe lower performance on *orphans & vulnerable children* and *u-report*, as the background knowledge provided for these classes is insufficient and not as good as the pat-

Dataset	Naïve Bayes	Pooling-AUC	Pooling
education	86.9	**87.7**	**88.1**
emergency	69.7	**71.5**	**72.1**
employment	77.8	**79.5**	**80.4**
energy	85.2	**85.6**	**86.4**
family	64.9	**67.7**	**68.0**
health	89.8	89.7	89.8
orphans	75.9	74.2	74.4
social	71.8	71.9	72.7
u-report	90.0	90.0	86.5
violence	75.4	**76.9**	**77.8**
water	94.0	94.1	94.2
Mean	80.1	80.8	81.0
Sig. W/D/L		6–4–1	6–3–2

Table 2: AUC performance of variants of Pooling Multinomials versus Naïve Bayes. Statistically significant improvements over Naïve Bayes are presented in bold.

terns discovered from the data. For instance, for *orphans & vulnerable children*, the most discriminative patterns we see from the labeled messages involve `early marriage, domestic violence,` and `child labour,` which are all missing in the keyword lists.

In the case of *health & nutrition*, we see that there are sufficient training examples to capture all the background knowledge; and as such, Naïve Bayes achieves the same performance as Pooling Multinomials for this dataset. An added advantage of using dual supervision is that one need not provide as many instance labels to achieve high performance, thus reducing the burden of labeling data. Learning curves (as in [5]) reveal that Pooling Multinomials can achieve the same accuracy as Naïve Bayes with far fewer training examples.

In comparing the variants of Pooling Multinomials, we find that the hand-picked weight of 0.5 performs a little better than automatically selecting a weight. This is because automated selection is based on AUCs evaluated on the training set. This tends to be over-optimistic for the model built on labeled instances versus labeled features, since the former is trained on the same data it's evaluated on. This issue may be alleviated to some extent by using cross-validated estimates on the training set. In general, automatically selecting weights is crucial in a setting where the number of labeled instances or features is changing – since the relative performance of the two components will be shifting. However, if you have a fixed dataset, and want to build the best classifier, we recommend hand-picking a weight based on tuning performance on a validation set.

6. MODEL RE-RANKING

In this section, we discuss the interpretation of AUC results, and present an approach for further improving the ranking produced by our models.

6.1 AUC in Practice

When there is a high class imbalance in the data, area under the ROC curve may present a more optimistic picture of model performance [2]. This is the case in our data, were the number of relevant messages is much smaller than the number of irrelevant messages for each class — on average only 3% of messages are relevant to a particular class. Consider the results of using a Naïve Bayes classifier for *water*, where we achieve an AUC of 94%. While this may seem high in absolute terms, an examination of the top 100 messages ordered by the posterior class probabilities reveals that only 46 of these messages actually pertain to water and sanition-related issues. The AUC for a class can be interpreted as the probability that a classifier will rank a randomly chosen relevant instance higher than a randomly chosen irrelevant one [3]. The reason we observe such as high values for AUC is that the background probability of a random message being relevant to *water* is only 1.9%. So a precision of 46% in the top 100 is indeed vastly better. However, since UNICEF employees manually inspect messages at the top of the list, in deployment, this is still an unacceptably high false-positive rate.

6.2 Re-ranking Classifier

In contrast to text classification, the advantage of using a keyword matching approach, as in Section 4.1, is that the top of lists tend to be high in precision. This is because the lists were constructed with high-precision patterns, such as, any mention of an `ebola outbreak` should be flagged as relevant to *emergency*. However, this comes at the high price of low recall, as the lists are far from comprehensive.

On the other hand, the supervised learning models tend to have higher AUCs, but do not guarantee high precision at the top. However, high-precision keyword lists can be directly leveraged to improve on model-based ranking. This can be done if we reorder the model rankings in a way to meet the following criteria:

- If a message matches at least one keyword, it should be ranked above messages that do not have a match (irrespective of model score).
- Amongst all messages that have a match, retain the ordering given by the model.
- All messages that do not have a match, appear at the bottom of the list, ranked by model score.

In essence, we want to retain the original model ordering, as far as possible, but have all messages that have a high-precision keyword match bubble up to the top of the list.

We can think of this *Re-ranking Classifier* as an ensemble of two classifiers:

- The first, is a *Keyword Classifier*, which, for each message, returns a posterior class probability distribution $[q, (1-q)]$, that can only take on two extreme values $[1, 0]$ and $[0, 1]$, depending on whether there is a match or not.[2]
- The second, is a text classification model (Naïve Bayes or Pooling Multinomials), which returns a posterior distribution $[p, (1-p)]$.

The posterior distribution of the ensemble is then computed based on a weighted sum of the two classifiers, i.e.,

$$[(\omega q + (1-\omega)p), 1 - (\omega q + (1-\omega)p)]$$

Now, the re-ranking criteria describe above will be satisfied for any $\omega > 0.5$. In practice, we use $\omega = 0.5 + 1e-6$, and refer to this classifier in our results as *Re-ranking*. Since the model performance is often significantly better than keyword classification, we also compare to using $\omega = 0.25$, which places a higher weight on model ranking, while still incorporating the Keyword Classifier.

[2] q is the probability the message is relevant to the target class.

6.3 Results

In order to test this approach, we take as our base text classifier the best performing classifier for each dataset from Table 2 (Naïve Bayes for *orphans & vulnerable children* and *u-report*, Pooling Multinomials for the rest), and apply Re-ranking to it. The results are summarized in Table 3, where we also include the baseline of using just keyword matching.

We observe that model re-ranking shows size-able improvements over the base text classification results. The best results are for Re-ranking ($\omega = 0.25$), where all improvements over the base classifier are statistically significant, and on average leads to a 5% increase in AUC.

In order to quantify the impact of our entire process of spell-correction, dual supervision and model re-ranking, we can compare to the original Keyword Classifier baseline (column 5 in Table 3). This highlights the dramatic impact of the modeling enhancements leading to improvements ranging from 2.4% to 39.5%, with an average increase of 19.8% in AUC.

We note that counter to expectation, Re-ranking ($\omega = 0.25$) performs as well as or better than Re-ranking ($\omega = 0.5 + \epsilon$), for most datasets. This is in part due to the fact that the UNICEF keyword lists are not all high-precision patterns. For example, the word `well` is on the list for *water*, but clearly not all messages using `well` are in the context of water. Further improvements can be made using a well-vetted keyword list limited only to high-precision patterns.

7. DEPLOYMENT AT UNICEF

In this section, we discuss some of the issues for consideration in deployment and present some qualitative results of the system in action.

7.1 Allocating Resources

Thus far we have used area under ROC curves to guide all our modeling choices. However, AUC provides a measure of performance on ranking of entire lists, and as discussed before, need not reflect the precision at the top of the lists. If we knew *a priori* that only the top k messages in a list would be examined, we can attempt to directly optimize a metric such as Precision@k [1, 4]. However, part of the challenge in this deployment is determining how to appropriately allocate resources for manual inspection of lists. UNICEF needs to ensure the most important messages are addressed by the human resources available. In the absence of knowing the resource allocation *a priori*, we find it best to use AUC to direct model choices.

However, in order to drive decisions on how to allocate human bandwidth, we look at plots of Recall (true-positive rate) versus number of messages examined. In particular, we order messages in descending order of models scores, assuming a user will inspect messages in this order. The plots for all four systems in Table 3 are presented for all 11 datasets in Figure 2. These plots help us determine the relative resources required for different datasets. For instance, just inspecting the first 2000 messages captures more than 90% of all the messages relevant to *water*. While for *employment*, we would need to examine 7000 messages to capture 80% of the true positives. The resource allocations done by UNICEF further factors in the importance of each class, which translates to different target recall rates – for instance, capturing all relevant messages is more important for *emergency* versus the *u-report* category.

The recall plots also demonstrate the difference between models at different operating points, not captured by AUC. For instance, for *u-report*, around 10,000 messages, Naïve Bayes has the best recall, but Re-ranking performs better as we go further done the list. Such plots can help us make more informed choices between models taking recall versus resources trade-offs into account.

Finally, these plots clearly illustrate the size-able impact of our modeling enhancements over the initial baseline of keyword matching. This improvement was critical in making it possible for UNICEF to rely on automated routing of messages.

7.2 Impact of Deployment

The output of our process is a ranked list of messages which are routed to the appropriate UNICEF departments. Table 4 shows a sample of the top 15 messages scored as relevant to *water* by our model. As a point of contrast, we also present the list of top messages from a baseline keyword-based process. Despite having a high AUC, 5 of the top 15 messages are actually irrelevant in the baseline system; while all of the top 15 are relevant in the model-scored list. While this is a small sample on a single dataset, these results are representative of the improvements in our message ranking system.

Figure 3 shows a screen-shot of the U-Report administrative user interface, where UNICEF employees in the appropriate department, can look through the message list ranked by the model, and note whether information needs to be provided or an action needs to be taken. They can also indicate the urgency of a message by providing a rating.

Once a message is classified as *informational*, it is usually responded to individually by experts online using the U-report UI. If there is a high volume of messages on a topic, then a mass campaign is enacted. For example, having received over 500 messages about nodding disease, UNICEF Uganda identified a demand for information in a specific region of Uganda. Subsequently they sent a series of mass text messages to everyone living in the affected region, instructing them how to recognize symptoms and seek treatment for the disease.

Messages identified as *actionable* are prioritized and validated by UNICEF experts, and shared with the relevant government agency for action. These messages generally constitute an emergency. For example, a typhoid outbreak detected was communicated to the Ministry of Health and District Health Officer, to understand the situation and to provide additional support as needed.

All 386 Members of Parliament in Uganda are now subscribed to U-report. Through this system they receive SMS updates on reports pertaining to key issues in their districts. This enables the MPs to stay abreast of the latest issues affecting their constituents, and to take parliamentary actions to address these issues.

By elevating the voice of youth to political levels, the U-report program has already resulted in notable policy changes in Uganda. For instance, the requirements for accessing a youth loan scheme, called the Youth Fund, were changed after U-reporters complained of the O-levels (high-school diploma) requirement for eligibility. This requirement was subsequently eliminated and the change was announced exclusively by the Ministry of Gender, Labor and Social Development on the U-report TV show.

Dataset	Base Classifier	Re-ranking ($\omega = 0.5 + \epsilon$)	Re-ranking ($\omega = 0.25$)	Keyword Classifier
education	88.1	**90.6**	**91.1**	79.8
emergency	72.1	**76.8**	**76.8**	60.8
employment	80.4	**89.7**	**89.4**	80.5
energy	86.4	**92.2**	**92.2**	**90.0**
family	68.0	**69.8**	**69.8**	55.3
health	89.8	**90.9**	**91.5**	77.5
orphans	75.9	**78.4**	**78.4**	56.4
social	72.7	**76.6**	**76.6**	58.8
u-report	90.0	87.0	**90.4**	64.8
violence	77.8	**83.3**	**83.3**	64.7
water	94.2	**98.4**	**98.2**	94.4
Mean	81.4	84.8	85.2	71.2
Sig. Win/Draw/Loss		10–0–1	11–0–0	1–2–8

Table 3: AUC performance of Re-ranking versus the base classifier and the Keyword Classifier baseline. Statistically significant improvements over the base classifier are presented in bold.

Another notable example of the impact of U-report is in the area of child abuse. Following a sustained effort to end corporal punishment in schools, the Ugandan parliament proposed the Children Act Amendment banning such abuse. This amendment has passed into the final stages of being written into Ugandan law. This change followed from a coordinated effort between the Chief Justice and U-reporters to highlight why beatings in schools have to end.

In order to drive change at a local government level, all Chief Administrative Officers (one for each of the 112 districts) have agreed to receive SMS updates on issues in their localities. To further assist in understanding regional issues, the classifications in our system are used to power a National Pulse website, which provides a visualization of the trending topics in development issues by geographic regions, based on the origins of the SMS messages.

Message	True Label
i don't know how we can end corruption in uganda we are at risk anon contracting water borne diseases in our village we drink and fetch water for domestic use from a spring water called Oz the top cover lid anon the water source has recently been opened to clean it and it's channel to the pipe but to cover it has become a problem the water committee collected money for buying cement and other necessities but instead diverted it into their various pockets	relevant
water is life water is a necessity but safe water is scarce in the region because wells are drying up and tap water is not regular leading to high costs anon water and living at large	relevant
in this village we have a problem anon water most bore holes have broken down we don't have piped water taps in fact Pele start lining up 4 water	relevant
in our village we fetch water down the hill on flowing river the hill is steep stony and very long and the source anon river is icu ya forest with baboons which dir ten the water so we have problem anon water mostly	relevant
ate central aw ila is one anon those villages that has suffered water shortages bad health conditions many Pele hv had to walk long distances yet not for safe but dirty water bc oz anon no option left even the boreholes hv rusted from inside give dirty brown water...	relevant
use clean safe water always drink boiled water to avoid diseases save water and don't allow animals to play or drink water you use at home don't put you	relevant
hi ureporters we've a problem anon water pollution by a certain coffee station it's called Kakalina in blamable they pour smelly and dirty water yet some people drink this guzzle stream water...	relevant
no proper water source in our Villa Pele are drinking very dirty water no water gard aqua safe place help	relevant
in my community we have problem anon water we don't have clean water 4 drinking we are using pipe water which is not saved 4 drinking	relevant
in my area people are facing a very big problem anon clean water source borehole is far hence drinking dirty water from a wel especially old mothers and they are infected with typhoid and ditherer	relevant
dear here in my villa gm there is acute shortage anon water where there is no bore hole people sink water in valley share with animals	relevant
dear here in my village there is acute shortage anon water where there is no bore hole people sink water in valley share with animals	relevant
there is a possible out break anon cholera in water village because anon poor sanitation worse still the water table is very high only meter's from the ground hence all latrines are shallow and with the onset anon the rains means a possible out break anon cholera or any water borne disease	relevant
in paw el angina people are drinking dirty water no any borehole 3oo Pele fetch stream	relevant
water shortage in the surrounding areas there is plenty anon gravity flow water but to my surprise water is given to far areas leaving this zone Cabot	relevant

Table 4: Top ranked messages relevant to *water*, using our deployed model.

Message	True Label
for water butwe are also sharing with animals.i want advise from u-rep	relevant
e here we are lacking clean water.we are moving long distance looking	relevant
how do u use hard water 4 washing	relevant
najja subcounty in buikwe district lacks clean water...	relevant
plize I'm not understanding vary well can u give me detail about ureport.	irrelevant
dirty water we've boreholes but siezed working due to mechanical problems we need assistance to repair them...	relevant
sible most especially spring wells & boreholes water have become	relevant
s at acost the public has however not been well informed about the h	irrelevant
2 men lost their lives in the pita <anon> latrine when they were digging the pit and their bodies was being removed in the morning.	irrelevant
i was not impressed to see someone in charge <anon> a bore hole in paibona where i visited increasing the fare for maintence from 300 to 500...	relevant
najja subcounty in buikwe district lacks clean water...	relevant
st homes have been flooded with water property worth millions...	irrelevant
residents <anon> aworanga are holding a meetin today urging gov'nt to construct safe water sources.	relevant
the fate <anon> students in rural communities <anon> uganda.(as soon as the last bell goes which signifies the end <anon> school <anon> collects her bks...	irrelevant
we are lacking clean drinking water in yapi	relevant

Table 5: Top ranked messages relevant to *water*, using a keyword baseline approach.

8. CONCLUDING REMARKS

In order to provide a voice to the youth in Uganda, we have deployed a high-precision SMS routing system, so that their concerns may be addressed by the appropriate authorities in a timely manner. In the process of developing this text classification system, we have learned a few lessons that may be broadly applicable to other domains.

First, auto-correction of text is indeed an effective preprocessing step for SMS messages. While simple to implement, and not necessary always accurate, this pre-processing step does make a significant impact on supervised text classifiers. This process could also be applied to other domains analyzing short messages, where authors abandon the orthography of formal writing, such as in Twitter.

Second, background domain knowledge should not be ignored in place of supervised leaning. With dual supervision we were able to go beyond standard text classification, by also leveraging known associations between words and classes in our domain. A small amount of time spent extracting such domain knowledge from experts (labeling words) can significantly reduce the amount of effort typically dedicated to only labeling training examples.

Third, high-precision keyword patterns can be systematically used to re-order supervised model scores, in order to produce a ranking that combines the best of both approaches.

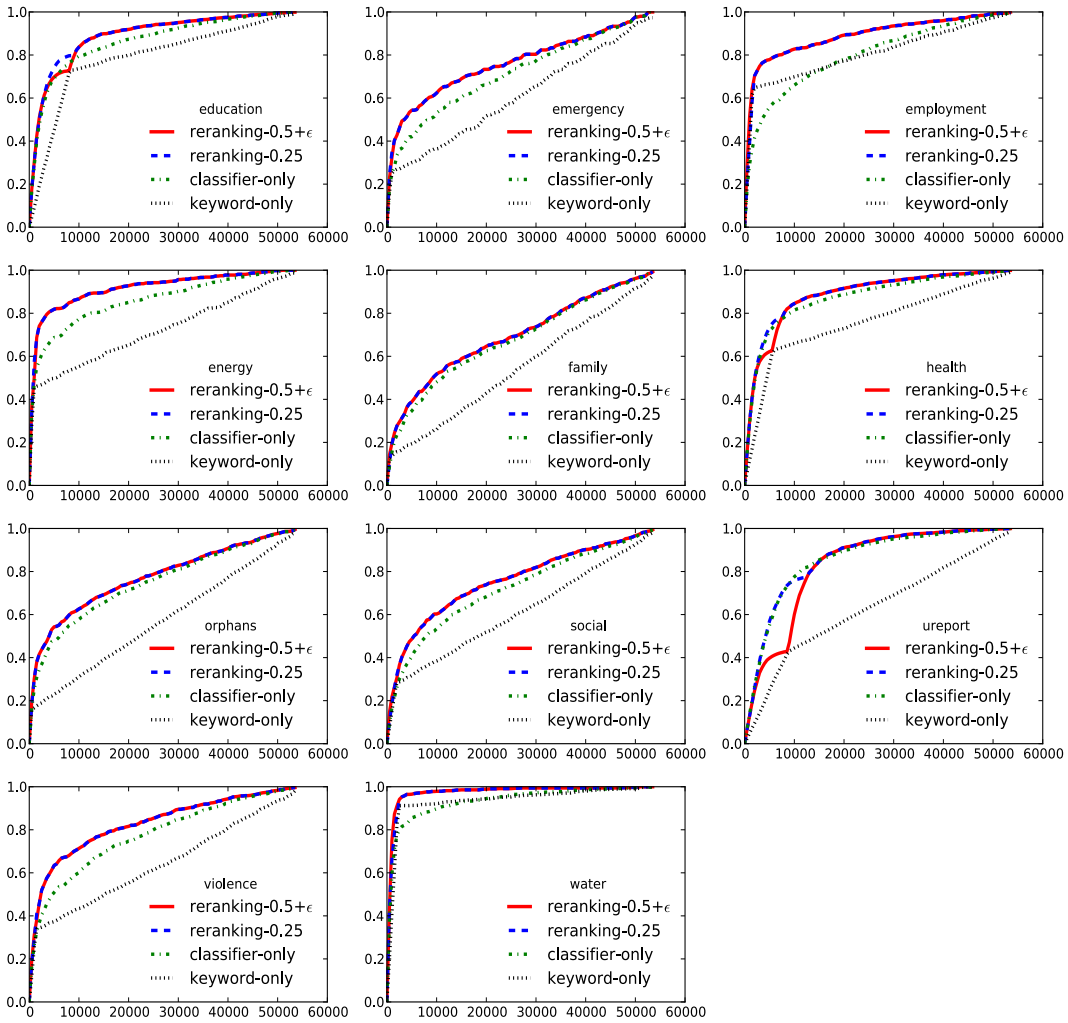

Figure 2: Recall (y-axis) vs. number of messages processed (x-axis) for different approaches.

A combination of the three steps above results in notable model enhancements in our domain, producing up to a 40% improvement in AUC over a keyword-based baseline. The resulting text analysis system enables the U-report program to effectively identify and address the concerns of young people in Uganda, on a range of social issues that impact their daily lives. Based on the success in Uganda, the U-report program has already been launched in Zambia, South Sudan, and Yemen, and will soon be deployed in the Democratic Republic of Congo, Zimbabwe and Burundi.

9. ACKNOWLEDGMENTS

We would like to thank Abraham Okiror for labeling data; and Phaneendra Divakaruni for help with the U-report UI. We would also like to acknowledge Joshua Harvey for making this deployment possible.

10. ADDITIONAL AUTHORS

Additional authors: Sharad Sapra (UNICEF Uganda, email: `ssapra@unicef.org`), Rajesh Anandan (US Fund for UNICEF, email: `ranandan@unicefusa.org`) and Solomon Assefa (IBM Research, email: `sassefa@us.ibm.com`).

11. REFERENCES

[1] C. J. C. Burges, K. M. Svore, P. N. Bennett, A. Pastusiak, and Q. Wu. Learning to rank using an ensemble of lambda-gradient models. *JMLR - Proceedings Track*, 14:25–35, 2011.

[2] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006.

[3] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[4] T. Joachims. A support vector method for multivariate performance measures. In *ICML*, 2005.

[5] P. Melville, W. Gryc, and R. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *KDD*, 2009.

[6] V. Sindhwani and P. Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *ICDM*, 2008.

[7] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, pages 412–420, 1997.

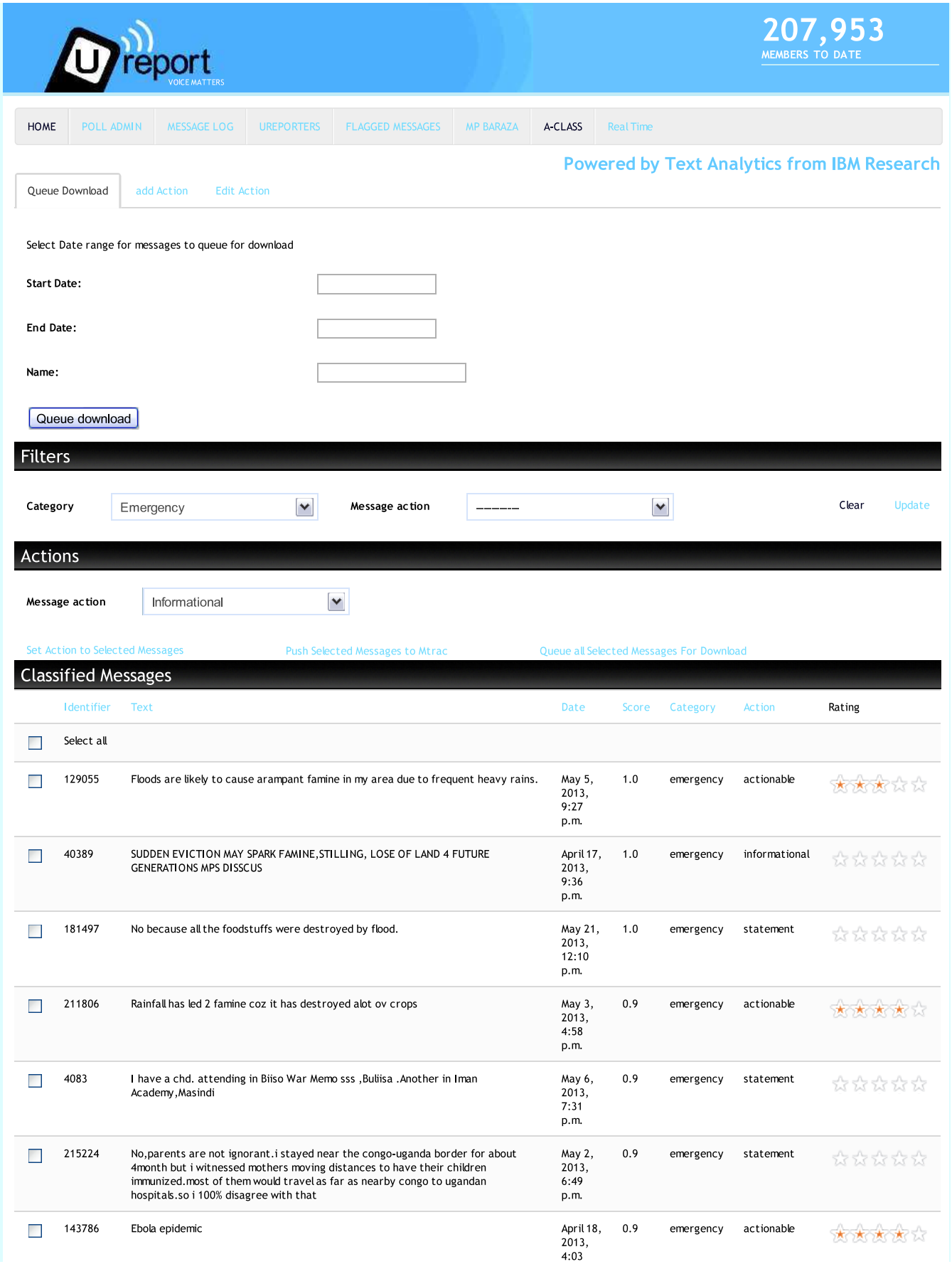

Figure 3: U-Report administrative user interface.