Classification models for new language communities: building domain-specific message categorization

Jessica Long
Idibon
jessica@idibon.com

Robert Munro
Idibon
rob@idibon.com

Nicholas Gaylord
Idibon
nick@idibon.com

Kidus Fisaha Asfaw
UNICEF
kasfaw@unicef.org

Evan Wheeler
UNICEF
ewheeler@unicef.org

ABSTRACT

U-Report is a social messaging tool developed by UNICEF that allows people in developing countries to sign up to respond to polls, report issues, and become agents of positive change within in their communities. This program, which has grown to serve over 15 countries in 3 years, is already outpacing the ability of UNICEF and its partners' human workforce to review incoming messages. Incoming messages vary in their urgency, from greetings and appreciation to imminent public health and security risks. In this paper, we present a procedure for automatically identifying and classifying urgent messages via bootstrapping human-labeled data to build a semi-supervised classification model.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Lang. Processing

General Terms

Algorithms, Measurement, Experimentation, Languages

Keywords

Crowdsourcing, machine learning, multilingualism, natural language processing, classification, implementaiton

1. INTRODUCTION

Semi-supervised classification is a well-known tool for text categorization, yet in many domains it remains difficult to use because of need for human-labeled training data. This can be difficult to obtain, particularly in low-resource languages [1, 2], including those served by U-Report. This is exacerbated by the fact that U-Report messages are sensitive and cannot be shared externally. Additionally, the data is dynamic: thousands of new U-Reporters join every week and the conversation shifts with each new poll question, which poses challenges for ongoing model tuning. We show how creating a feedback loop between human judgments and machine predictions both accelerates annotation and improves message classification accuracy.

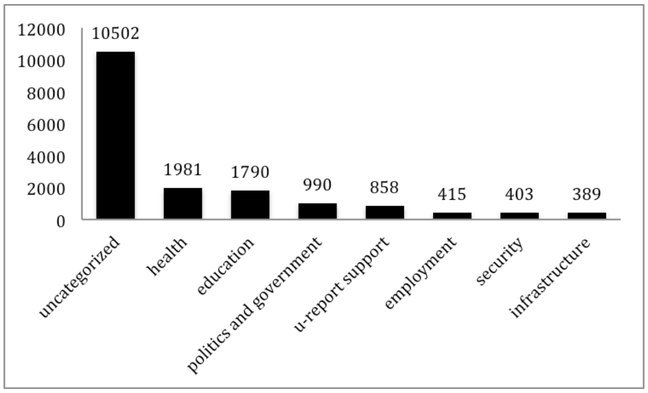

Figure 1: Frequency of labels in 5 days of data

2. CLASSIFICATION MODEL BUILDING

2.1 Bootstrapping Human Labels

U-Report has two main sources of data: **Poll responses** (multiple-choice or open-ended messages tied to a given poll), and **Inbox Messages** (open-ended and often unsolicited).

Our goal was to sort messages into the following seven categories: Health, Education, Politics & Government, U-Report Support, Infrastructure, Employment, and Environment. These categories are coarser-grained than the topics of individual poll questions, and any individual poll may be relevant to more than one category. However, we were still able to utilize pre-existing open-ended poll responses to help train classifiers for these coarser-grained category labels.

For example, from thousands of responses to a poll about children's safety, we took the 487 longest responses and identified 357 messages (73%) that were related to "Security". Using poll data in this fashion means that a training set can be amassed very quickly. To find 3 examples about "Security" in the constructed dataset, a human must review only 4 messages, versus 150 in the data overall.

We supplemented this data with additional annotations. We annotated a sample from the top 20% longest messages in the U-Report corpus, and found that only 11% of labels on this data match the most frequently occurring cate-

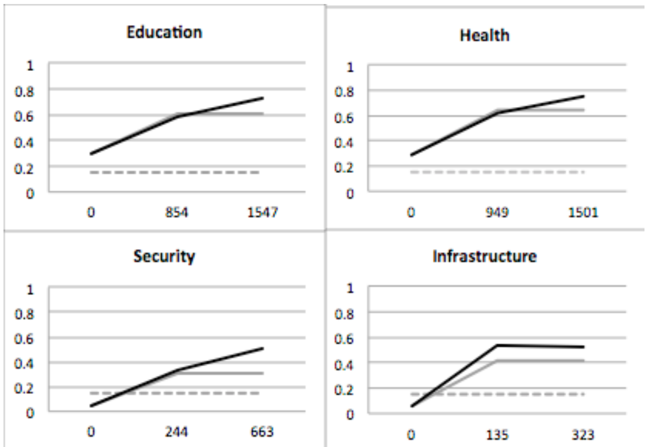

Figure 2: Classification f-score against baselines.

gory, "health." This underscores the value of using partially-related data to bootstrap building a labeled training set. Figure 1 shows the distribution of labels in a fully labeled dataset over five days.

2.2 Model training and testing

We used the above data to train seven independent classifiers, one for each category, to enable a given message to be associated with more than one category in cases of subject matter overlap. Best performance was achieved using unigram features only. Figure 2 shows f-score as a function of annotation count on 4 labels. Classification performance with zero labels assumes uniform positive prediction.

We compare our model performance to two baselines. First, we assume that UNICEF analysts can review 400 messages per day, and that their precision is equal to the observed inter-annotator percent agreement on their dataset (91%). This baseline measure assumes classification accuracy at that rate on those 400 messages, and no classification on the remaining messages per day. (we assume a constant number of 5000 messages per day, or roughly the current rate). This value is plotted as a dotted gray line. Our second baseline is based on a model using hand-tuned unigram features selected from a list of keywords generated from a topic model [3]. This baseline is plotted a solid gray line.

Because of our a priori knowledge, the hand-built model initially performs comparably to ours. However, the limitations of manual feature selection eventually become evident. Table 1 shows some of the top predictive features from our model. Each bolded feature pair indicates a feature that was not chosen for the manual model. The least intuitive feature, "true", comes from a time when rumors and misinformation were propagating about the Ebola crisis, with messages containing "is it true that..." referencing various rumors about preventative measures (including bathing in *salt water*, another bolded feature in Table 1).

The distribution of message labels underscores the importance of this work. While "Security" is one of the sparsest labels, automatically identifying messages where respondents express fear for their safety is one of the most important capabilities of a classification system. These messages are relatively infrequent, but are very often urgent when they do occur. Messages such as these are at greatest risk of being overlooked by the baseline strategies we compare against, particularly the 400 message/day baseline.

Category	Feature	Category	Feature
health	ebola	**health**	**salt**
u-report support	report	education	school
health	virus	u-report support	register
employment	paid	health	water
health	**true**	infrastructure	water
u-report support	reporter	employment	job
health	disease	**health**	**contact**

Table 1: Top features from our unigram model.

2.3 Results and Implementation

We saw significant increases in f-score from the naïve approach to ours for all classifiers we trained. The gains ranged between 0.16 for employment to 0.6 for health. We currently label thousands of messages per day that U-Report Nigeria receives in real time. We include a holdout set of 5% of messages close to our classification decision boundary, as well as 5% of messages uniformly sampled from across the set of unpredicted messages. These holdout sets allow us to continue monitoring precision and recall, respectively, and can be used to further iterate semi-supervised training of our predictive models.

3. CONCLUSION

Layering structured understanding on top of unstructured text is a critical capability as more of the world comes online. This is particularly evident in the U-Report dataset, where the diction and structure of Nigerian writing differs significantly from existing labeled data.

Applying existing semi-supervised classification algorithms relies on the ability to effectively and efficiently solicit human labels for messages. We took advantage of existing implicitely labeled data in a form that was very similar to what we eventually hoped to classify. We hope this work can inspire future implementers who worry that they don't have the existing structured information they need to apply classic machine learning techniques.

Future work involves building classifiers from poll responses that will automatically allow the building of models based on thousands of implicitly labeled examples that can be collected in a few hours. We also plan to extend our language-agnostic text classification methodologies to build categorization models in additional non-English languages served by U-Report.

4. REFERENCES

[1] N. Gaylord, A. Palmer, and E. Ponvert, editors. *Proceedings of TLSX: Computational linguistics for less-studied languages*. CSLI Publications, 2008.

[2] R. Munro. *Processing short message communications in low-resource languages*. PhD thesis, Stanford, 2012.

[3] M. Steyvers and T. Griffiths. Probablistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Erlbaum, 2007.