# Detection of Planned Protests

Author List TBD
Somewhere, Sometime
some1@somewhere.edu

Author
Somewhere
someone@somewhere.edu

## ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings. It is an *alternate* style which produces a *tighter-looking* paper and was designed in response to concerns expressed, by authors, over page-budgets. It complements the document *Author's (Alternate) Guide to Preparing ACM SIG Proceedings Using LaTeX2$_\epsilon$ and BibTeX*. This source file has been written with the intention of being compiled under LaTeX2$_\epsilon$ and BibTeX.

The developers have tried to include every imaginable sort of "bells and whistles", such as a subtitle, footnotes on title, subtitle and authors, as well as in the text, and every optional component (e.g. Acknowledgments, Additional Authors, Appendices), not to mention examples of equations, theorems, tables and figures.

To make best use of this sample document, run it through LaTeX and BibTeX, and compare this source code with the printed output produced by the dvi file. A compiled PDF version is available on the web page to help you with the 'look and feel'.[1]

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

## General Terms

Theory

## Keywords

ACM proceedings, LaTeX, text tagging

## 1. INTRODUCTION

The *proceedings* are the records of a conference. ACM seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes (for instance, 9 point for body copy), a specified live area (18 × 23.5 cm [7" × 9.25"]) centered on the page, specified size of margins (1.9 cm [0.75"]) top, (2.54 cm [1"]) bottom and (1.9 cm [.75"]) left and right; specified column width (8.45 cm [3.33"]) and gutter size (.83 cm [.33"]).

The good news is, with only a handful of manual settings[1], the LaTeX document class file handles all of this for you.

The remainder of this document is concerned with showing, in the context of an "actual" document, the LaTeX commands specifically available for denoting the structure of a proceedings paper, rather than with giving rigorous descriptions or explanations of such commands.

## 2. RELATED WORKS

- **HRL's Twitter Planned Protest Paper: Detecting future social unrest in unprocessed Twitter**:

  *keywords*: 335 keywords identified by domain experts, used to filter twitter stream.

  *Date*: Quoting "Our future dates is done."Our future date filter searches first for month names and abbreviations in Spanish and Portuguese and second for numbers less than 31 within three whitespace separated tokens from each other. Thus, an example matching date pattern would be âĂİ10 de eneroâĂİ.

  *Probability*: Posterior probability of tweet being Civil Unrest related Given User-type

  *Event Geocoding*: If no location in Text then location of retweeter is used based on HRL's Geocoding methodology.

  *EventType and Population*: Duplicate Forecasts for same date/location are combined into one and then tweet history of every tweeter/retweeter in the forecast are searched for keywords pertaining to particular classes identified by domain expert and then most commonly occuring class is assigned.

  *Results Provided*: Counts of Warnings provided and results of Manual examination of warning provided

[1]Two of these, the `\numberofauthors` and `\alignauthor` commands, you have already used; another, `\balancecolumns`, will be used in your very last run of LaTeX to ensure balanced column heights on the last page.

- **HRL's Tumblr Based Paper- Civil Unrest Prediction: A Tumblr-Based Exploration**

  *keywords*: 59 keywords by domain experts.

  *location*: Text Based filter for pre-defined locations(1022 in number)

  *date*: Similar to the Twitter paper

  *Dataset*: Full tumblr firehose of public post from 2013-04 to 2013-11. Evaluations from 2013-06 to 2013-08. Results presented in terms of Number Events Detected, Precision, Leadtime

- **Identifying Content for Planned Events Across Social Media Sites - Luis Gravano-Columbia University**: Not exactly related to ours. It deals with identifying user contributed content for future planned events that are already known like music concert etc.

# 3. PROBLEM

# 4. APPROACH

## 4.1 Enrichment

- **Basis Enrichment**: We make use of the Basis Rosette Language Processing tools to identify text language (spanish, portuguese etc.,), tokenize, lemmatize and identify named entities as well as classify them into Locations, Person and Organization.

- **TIMEN**: Date expressions are normalized and de-indexed using the TIMEN package – *TIMEN: An Open Temporal Expression Normalisation Resource.*

## 4.2 Learning of Phrases

[Sathappan writes: *Intro written for KDD paper*] Initially, a few seed phrases were obtained manually with the help of subject matter experts. These phrases were parsed using a dependency parser and the grammatical relationship between the core subject word—*protest*, *manifestaciÃşn*, *Huelga*, etc.—and any accompanying word – *plan*, *call*, *anunciar* — was extracted. To extend the initial set of phrases, a set of tweets containing a subject word and a future time/date expression was collected and parsed. This set of sentences was used to expand the set of planned protest phrases by extracting all keyword combinations that have the same grammatical relation with respect to the core subject word. The final set of planned protest phrases is then obtained after a manual revision of the phrases obtained in the last step.

[Sathappan writes: *picture from CA-Tech Interview slides*]

## 4.3 Geo-Coding

We make use of different geocoding methodologies for geocoding news/blogs and twitter.

### 4.3.1 News/Blogs

Most news articles and blog posts mention multiple locations, e.g., the location of reporting, the location of the incident, and locations corresponding to the hometown of the newspaper. We developed a probabilistic reasoning engine using probabilistic soft logic (PSL) to infer the most likely city, state and country which is the main geographic focus the article.The PSL geocoder combines various types of evidence, such as named entities such as locations, persons, and organizations identified by RLP, as well as common names and aliases and populations of known locations. These diverse types of evidence are used in weighted rules that prioritize their influence on the PSL model's location prediction. For example, extracted location tokens are strong indicators of the content location of an article, while organization and person names containing location names are weaker but still informative signals; the rules corresponding to these evidence types are weighted accordingly.

The methodology is similar to *Web-a-where: Geo-Tagging Web Content.*

### 4.3.2 Twitter

The Twitter geocoding is achieved by first considering the most reliable but least available source, viz. geotags, which give us exact geographic locations that can be reverse geocoded into place names. Second, we consider Twitter places and use place names present in these fields to geocode the place names into geographical coordinates. Finally, we consider the text fields contained in the user profile (location, description) as well as the tweet text itself to find mentions of relevant locations which can then be geocoded into geographical coordinates.

### 4.3.3 Facebook

We make use of only the facebook event data for our experiments. Almost, all of the Facebook event pages contain information about the venue of the event which includes latitude, longitude, country, state, city, street etc. Under cases where only latitude and longitude is given we do a reverse-geocoding by a KD-Tree lookup from the World Gazetteer to get the country, state, city information. A very few event pages do not have sufficient location info, we ignore such pages that do not even contain country information.

## 4.4 classification

[Sathappan writes: *Is this Needed? ..We have two classifiers – Text Based Naive Bayes Classifier for RSS and Facebook. Location Based Classifier for Twitter*]

# 5. EXPERIMENTS

# 6. DISCUSSION

# 7. REFERENCES

[1] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.