# Social Unrest: Classification and Modeling

Dan Saadati, Farah Uraizee, Tariq Patanam

## Motivation

As social media rapidly becomes a podium for political opinions and a tool for the organization and facilitation of protests, a powerful stream of data documenting opinions and actions of individuals becomes readily available. This type of information can provide key social insights in predicting areas at risk of social unrest, which can be significantly useful in scenarios prone to violence.



## Data

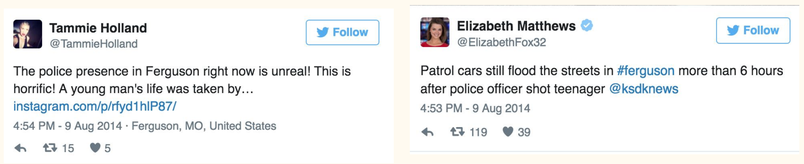


**Figure 1:** Tweets from sample social unrest event (Ferguson)

- Tweets queried with **location** and **time range**; each collection treated as **event**
- For training, used known events like Ferguson and Baltimore protests

## Preprocessing

Tweet Sanitization

- Normalized tweet format – lower-case, no punctuation, #'s removed

Stop Word Removal

- Removing from published minimal word list ('the', 'a', etc.) + Twitter-specific generated words (@username, #, 'RT')

## Methodology

### Baseline

SVM + Presence of Each Word:

- Lacked diversity in training data, used entire vocabulary as feature set
- Results: overfitted, very inaccurate

### Algorithm

1. Prepare data for feature extraction
2. **Background subtraction**: downweight words common to both unrest and rest situations; $f_D(w) = f_1(w) - f_2(w)$ where $f_1(w)$ is the frequency of a word *w* in social unrest tweets

**Bag of Words Approach**

Pick top *n* words with a frequency > *k*

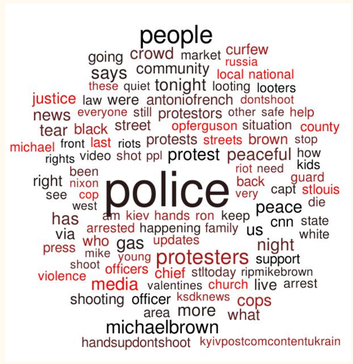


**Figure 3:** Most popular words in social unrest

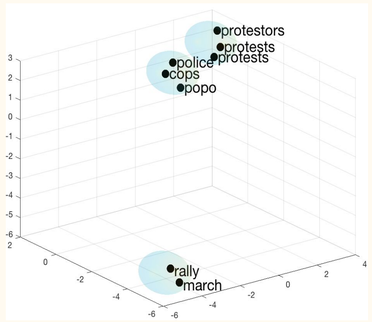


**Figure 4:** Clustering semantically similar words

### Bag of Clusters Approach

1. Get a vector representation of a word using **word2Vec**
2. Form word clusters using k-means and finding optimal *k* with loss function

$$J_\theta = Loss_{k-means} + Loss_k$$

$$Loss_k = \frac{1}{K}(Err(K=1) - Err(K=|V|))$$

3. These clusters represent similar words
4. Pick top *n* clusters with most word counts

## Results & Analysis

### SVM Regression Results

| | Baseline | Bag of Words (No BG Subtraction) | Bag of Words (BG Subtraction), America | Bag of Words (BG Subtraction), Global |
|---|---|---|---|---|
| Precision | 0.0 | 1.000 | 0.905 | 0.662 |
| Recall | 0.0 | 0.614 | 0.988 | 0.984 |
| F1 | 0.0 | 0.761 | 0.945 | 0.792 |

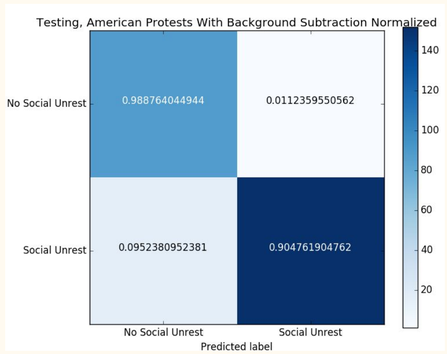


**Figure 2:** Confusion matrix for classifying American social unrest

**Results for American Protest**

| | |
|---|---|
| Training Error | 0.0259 |
| Testing Error | 0.2380 |

## Considerations

- Feature set overfits regionally
- Training Word2Vec model with Twitter data to more accurately reflect sentiment similarity

## Future Work

- **Linear Kernel SVM** - less prone to overfitting
- **Doc2Vec** - maps similar sentiment sentences to similar feature vectors
- **Tf-idf** - finds frequent terms across documents to help determine which words are relevant