

# Opinionizer Framework

## A Tool for Twitter Sentiment Analysis

Silvio Amir, João Filgueiras, Mário J. Silva and Bruno Martins

{samir, jfilgueiras, mjs}@inesc.id.pt, bruno.g.martins@ist.utl.pt



### Motivation

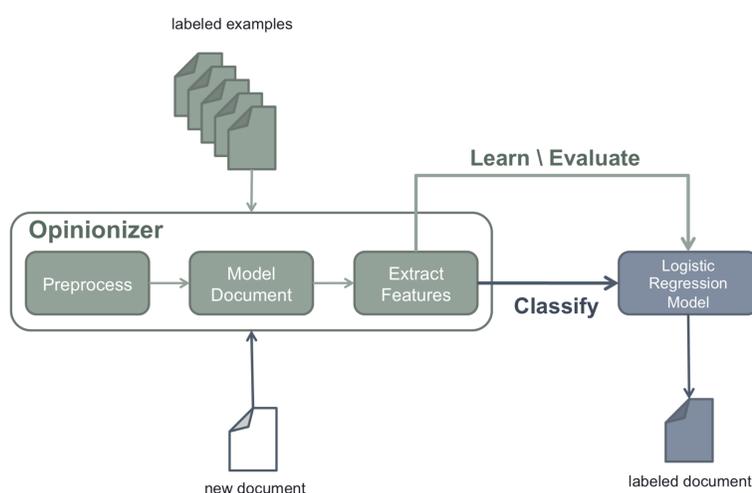
The rise of social media web applications, such as social networking sites, blogs and microblogs is revolutionizing the shape and behaviors of contemporary societies. It changed the way people access information and interact with each other providing an opportunity for:

1. the creation of a vast amount “crowd wisdom” that could be analyzed to infer characteristics of the population in real-time and make predictions about human-events.
2. the development of systems to exploit social media in order to measure public opinion, evaluate the popularity of products and brands, anticipate stock-market trends or predict electoral results.

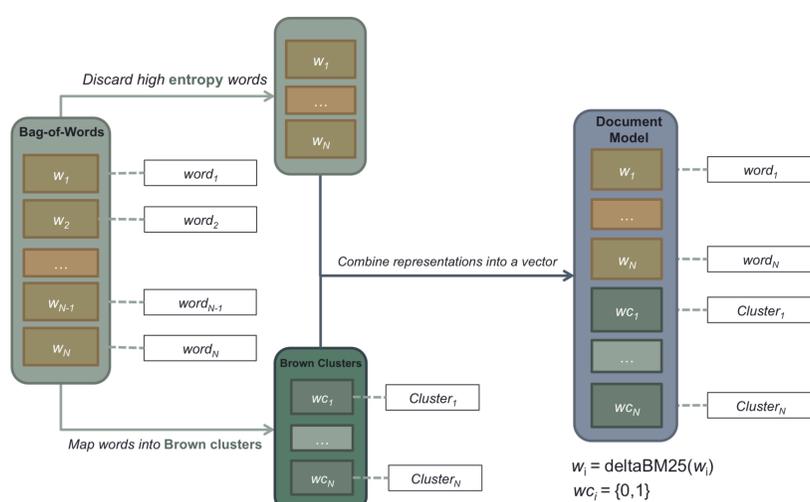
### Challenges of Twitter Sentiment Analysis

- **Noisy text**
- **Internet and microblog specific language**
- **Lack of context**, due to short messages
- **Sparse feature vectors**, also due to short messages and high lexical variation introduced by typos, abbreviations and “creative spelling”

### Proposed Approach



### Refined Document Model



Cluster	Sample words
0011000	gonna gunna gona gna guna ganna qonna gonnna gana qunna gonne goona
111010100010	lmao lmfao lmaoo lmaooo hahahahaha lool ctfu rofl loool lmfao lmfaooo lmaoooo lmba lololol
111101011000	facebook fb itunes myspace skype ebay tumblr bbm flickr aim msn netflix pandora
11101011011100	smh jk #fail #random #fact smfh #smh #winning #realtalk smdh #dead #justsaying

Figure 1: Brown clusters sample

$$\text{deltaBM25}(w_i) = t_{f_i} \cdot \log \frac{(N_{pos} - df_{i,pos} + 0.5) \cdot df_{i,neg} + 0.5}{(N_{neg} - df_{i,neg} + 0.5) \cdot df_{i,pos} + 0.5} \quad (1)$$

Figure 2: deltaBM25 -  $t_{f_i}$ : term frequency;  $N$ : document collection size;  $df_{i,c}$ : document frequency for class  $c$

$$\text{entropy}(w) = - \sum_{i=1}^N p(w|C_i) \log p(w|C_i) \quad (2)$$

Figure 3: entropy -  $N$ : number of classes ( $N = 3$  in this task);  $C_i$ : class  $i$

### Features

The document model was enriched with the following groups of features :

- **Negation**: words between a negation word and the first punctuation mark were suffixed with the `_NEG` token.
- **Sentiment Lexicons**: frequency of words with positive/negative prior polarity and a score obtained by summing both. Negation was taken into account by detecting the presence of a negation token in a window of two words. Words were extracted from Bing Liu’s Opinion Lexicon and SentiStrength internal sentiment lexicon.
- **Syntactic Patterns**: features that aim at capturing the creative and informal use of language in tweets, such as the number of sequences of exclamation marks or question marks, presence of emoticons, upper-cased words and emphasized words (with repeated characters).

### Evaluation

The Opinionizer was evaluated in tasks proposed by two competitions related to the field of Twitter Sentiment Analysis:

**SemEval 2013 - Twitter Sentiment Analysis** task: “Given a message, decide whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen”

**RepLab 2013 - Polarity Classification** subtask: “Develop systems dedicated to monitor the reputation of entities (companies, organizations, celebrities, etc.) on Twitter. (...) Given a message decide if the content has positive or negative implications for the entity’s reputation”

The framework was also used in the scope of **POPSTAR**<sup>1</sup>, a research project that aims at studying the use of social media to measure public opinion. We present the experiments for the development of a classifier “capable of measuring the sentiments expressed regarding parties, political actors and the economy in the contents of the Portuguese Twitter community”.

### Tuning

To optimize the classification performance and deal with fact that Twitter data samples can have a skewed class distribution, for each application:

- Feature selection is performed
- Logistic Regression hyper-parameter and cost function are tuned

### Experiments

We performed experiments with cross-fold validation on training data and measured Accuracy,  $F_1$  average and Polar  $F_1$  average (positive and negative class). The baseline consisted of a classifier trained with a binary weighted bag-of-words model with negation detection. Table 1 shows the results obtained in each step of our approach.

	SemEval			RepLab			POPSTAR		
	A	$F_1$	Polar $F_1$	A	$F_1$	Polar $F_1$	A	$F_1$	Polar $F_1$
Baseline	0.68	0.62	0.56	0.73	0.66	0.7	0.81	0.55	0.42
delta-BM25	0.72	0.66	0.62	0.73	0.66	0.71	0.84	0.57	0.44
Word Entropy	0.74	0.69	0.65	0.74	0.68	0.72	0.85	0.62	0.51
Brown Clusters	0.75	0.70	0.66	0.75	0.68	0.72	-	-	-
Features	0.75	0.72	0.69	0.75	0.68	0.73	0.85	0.62	0.51
Tuning	0.75	0.72	0.71	0.75	0.68	0.73	0.85	0.71	0.64

Table 1: Experimental results

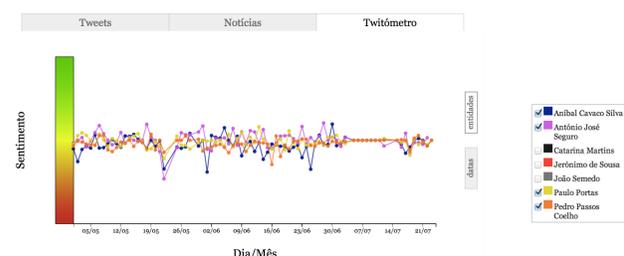


Figure 4: Opinionizer in POPSTAR - sentiment expressed towards political leaders in Portugal

### Conclusions and Future Work

We addressed the main challenges of performing sentiment analysis in the context of Twitter and built a tool that achieves state-of-the-art results in several related tasks.

Using this framework we participated in the RepLab 2013 competition and ranked among the top 3 for all the evaluated metrics, showing the effectiveness of the approach.

As future work we will induce Brown clusters and a sentiment lexicon from a corpus of Portuguese tweets. We also intend to further address the problem of unbalanced datasets with stacking or ensemble classification strategies.

This work was partially supported by FCT (Portuguese research funding agency) under project grants UTA-Est/MAI/0006/2009 (REACTION) and PTDC/CPI-CPO/116888/2010 (POPSTAR). FCT also supported scholarship SFRH/BD/89020/2012. This research was also funded by FCT under contract Pest-OE/EE/LA0021/2013.

<sup>1</sup><http://dmir.inesc-id.pt/projects/POPSTAR>

