Detecting Controversial Events from Twitter

Ana-Maria Popescu
Yahoo! Labs
Sunnyvale, CA
amp@yahoo-inc.com

Marco Pennacchiotti
Yahoo! Labs
Sunnyvale, CA
pennac@yahoo-inc.com

ABSTRACT

Social media provides researchers with continuously updated information about developments of interest to large audiences. This paper addresses the task of identifying controversial events using Twitter as a starting point: we propose 3 models for this task and report encouraging initial results.

Categories and Subject Descriptors

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

General Terms

Algorithms

1. INTRODUCTION

The explosion of social media has allowed researchers unprecedented access to data about the opinions of large audiences regarding political developments or popular culture events. The automatic detection of events which engage large social media audiences is a tempting challenge from both a sociological and a practical perspective: for example, displaying engaging news and events would allow web content providers to draw more users to their sites.

We present methods for detecting a specific type of engaging events - *controversial events* - using social media as a starting point. *Controversial events* provoke a public discussion in which audience members express opposing opinions, surprise or disbelief. Examples include incidents which violate the public's expectations about a particular entity, or which go against established social norms (see Table 1).

First, we introduce the notion of a Twitter *snapshot*, i.e. a triple consisting of a *target entity* (e.g., Barack Obama), a given *time period* (e.g., 1 day) and *a set of tweets* about the entity from the target time period. Given a set of Twitter snapshots, controversial event detection can be modeled as follows: (i) assigning a *controversy score* to each snapshot (ii) ranking the snapshots according to the controversy score.

In practice, we focus on controversies involving *celebrities* (actors, politicians, etc.) from Twitter, but our model can be easily generalized to other types of micro-blogging sites and entity classes (e.g., organizations).

Entity	Date	Example Controversy
Barack Obama	10/09/2009	President Obama is awarded the Nobel Peace Prize
Harry Reid	18/11/2009	Majority leader Harry Reid unveils Senate version of the Healthcare Bill
David Copperfield	13/01/2010	Rape charges against David Copperfield are dropped

Table 1: Controversial events from Twitter.

The main contributions of this paper are the following:

(a) We formalize the task of 'controversial event detection' and introduce 3 regression machine learning models to address it; (b) We describe a rich feature set for the target task; (c) We report encouraging experimental results: our models register statistically significant performance increases over all baselines, including relevant previous work.

Related Work Our work draws on and advances a rich body of *opinion mining* and *event mining* research. The work closest to ours is that of Tsytsarau et al. [7], who propose an unsupervised approach for mining contradictions at different levels of time granularity from postings on a topic. They introduce a new *contradiction* measure which we incorporate as a feature; subsequently, we show that our models outperform a supervised version of their (unsupervised) approach. Another relevant paper, [4], studies weblog comments and the mining of mixed-sentiment threads - our task and constraints differ, but some features are similar (e.g., polarity information and user engagement features). In recent years, *event mining* has moved from news collections to social streams: e.g., [9] define events as sets of relations between social actors on specific topics over certain time periods - in contrast, we focus on events centered on a given entity in a 1-day period and use a supervised approach. [6] uses community detection methods over a keyword graph to discover events - however, no evaluation is conducted, while we report detailed experimental results.

2. CONTROVERSIAL EVENT DETECTION

In this work we use the following definitions.

Event. Given a particular target entity, an event is defined as an activity or action with a clear, finite duration in which the target entity plays a key role.

Controversial event. An event is controversial if it provokes a public discussion in which audience members express

CIKM'10, October 25–29, 2010, Toronto, Ontario, Canada.

opposing opinions or disbelief (rather than no reaction, or an overwhelmingly positive - or negative - reaction).

Snapshot. A snapshot is defined as a triple: $s = (e, \Delta t, tweets)$ where: e is the target entity, i.e. any type of concept or named entity (e.g. 'Barack Obama') ; Δt is a time period (e.g. one day); $tweets$ is the set of all tweets from the target time period which refer to the target entity.

Event snapshot. A snapshot describing one specific event [1]. This can be either a *controversial-event snapshot*, describing a controversial event, or a *non-controversial event snapshot*, describing a non-controversial event.

Non-event snapshot. A snapshot which does not describe any event (spam, generic discussions, etc.).

Controversial event detection. Given a set E of target entities and a set of snapshots $S = \{(e, \Delta t, tweets) | e \in E\}$, the task is to rank snapshots in S according to a controversy detection function that assigns a controversy score $cont(s)$ to each snapshot s in S. The function should assign higher scores to controversial-event snapshots and lower scores to non-controversial-event and non-event snapshots. The task can be decomposed into two steps: *Event detection* (separating event snapshots from non-event snapshots) and *Controversy detection* (ranking event snapshots obtained from the previous step according to $cont(s)$).

2.1 Detection Models

We model the controversial event detection task as a supervised machine learning (ML) problem, where each snapshot s is represented by a feature vector constructed from Twitter and other sources. We operationally define the components of a snapshot $s = (e, \Delta t, tweets)$ as follows: e = any entity contained in a list of about 100K celebrities (see Section 3); Δt = 1-day period; $tweets$ = the set of tweets in the time period mentioning the entity. We use Gradient Boosted Decision Trees (GBDT, fully described in [2]) as the ML framework. We propose the following models for the task:

Direct model, estimating the controversy score $cont(s)$ in a single step using a single ML regression model. The manually annotated training set is composed of positive examples (controversial-event snapshots), and negative examples (non-controversial-event and non-event snapshots).

Two-step pipeline model, using the two-step decomposition presented above. The *event detection* classification model selects event snapshots from the set S. The *controversy detection* regression model assesses the level of controversy $cont(s)$ for snapshots selected in the first step.

Two-step blended model, a *soft* variant of the pipeline model. The result of the *event detection* step is used for providing more evidence for the *controversy detection* regression model. Specifically, the controversy detection model takes as input the full set of snapshots S and employs as an additional feature the prediction confidence score returned by the *event detection* model for each snapshot.

We employ a variety of resources to derive a large set of features for our models. A 7,590 word **sentiment lexicon** includes positive and negative polarity words from OpinionFinder 1.5 [8] as well as more informal opinion terms mined from user reviews (as in [5]); each sentiment word w has an associated polarity strength score $sent(w)$. A **controversy lexicon** contains 750 controversial words derived from Wikipedia pages of people mentioned in the Wikipedia *controversial topic* list. A **bad words lexicon** of 687 English bad words was downloaded from the Web [2]. Finally, we use an **English dictionary** of about 100K part-of-speech tagged English words, obtained as output of the Brill tagger [1] trained over the Wall Street Journal and Brown corpora.

[1]Snapshots describing multiple events may also exist, though they are very rare.

Model	AvgPrec	Aroc
base-rnd	0.202	0.480
base-hst	0.243	$0.528^{\dagger}$
base-tsy	$0.290^{\ddagger}$	$0.582^{\ddagger}$
direct	$0.568^{\S}$	$0.824^{\S}$
pipeline	$0.540^{\S}$	$0.805^{\S}$
blended	$\mathbf{0.618}^{\S}$	$\mathbf{0.850}^{\S}$

Table 4: 10-folds experimental results. † indicates statistical significance at 0.95 level wrt *base-rnd*; ‡ wrt to *base-rnd* and *base-hst*; § wrt to all baselines.

Table 2 contains our features, organized as follows:

Twitter-based features capture snapshots' linguistic properties (TW-LING), structural and social graph information (TW-STRC), the intensity of the discussion about the entity (TW-BUZZ), the distribution of sentiment words in the snapshot (TW-SENT), and the level of controversy (TW-CONT). Sentiment features are derived by computing the polarity of each snapshot's tweet t: $pol(t) = \sum_{w \in t} sent(w)$, where $sent(w)$ comes from sentiment lexicon.

News buzz features (EX-BUZZ) capture the intuition that if an entity is buzzy in news articles at the same time it is buzzy in a Twitter snapshot, then the snapshot is likely to refer to a real-world event. To compute such features, we *align* news articles with the set of snapshot tweets: given a snapshot with Δt, we extract news articles issued in period $(\Delta t - 1, \Delta t + 1)$ in which the target entity is a *salient entity*, i.e. it is mentioned in the article headline, or it is one of the 3 most frequently mentioned named entities in the article.[3]

Web and news controversy features (EX-CONT) assess the past and present levels of controversy surrounding the target entity in the snapshot.

3. EXPERIMENTAL EVALUATION

We experiment using a set of 104,713 "celebrities" obtained as follows : Wikipedia category lists for Actors, Musicians, Politicians and Athletes are scraped and entities whose names have less than 3 characters are removed.

Gold standard. We collect 738,045 Twitter snapshots referring to any of the above entities, from a (July 2009 - February 2010) firehose. We remove snapshots which have less than 10 tweets, more than 80% non-English tweets and tweets with more than 80% overlapping tokens [4], obtaining 73,368 snapshots. The *gold standard* contains 800 randomly sampled snapshots labeled by two expert editors: 475 are

[2]http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/ and http://www.noswearing.com/dictionary.

[3]Our corpus collects articles from Yahoo! News 2009/2010.

[4]We verified by manual inspection on a small set of discarded snapshots, that about 97% of them are indeed irrelevant.

Family	*Type*	*Features*
Twitter-based features		
Linguistic	TW-LING-NOU	Percentage of tokens that are nouns (PoS come from the English dictionary in Sec. 2)
	TW-LING-VRB	Percentage of tokens that are verbs.
	TW-LING-BAD	Percentage of tokens that are *bad words* (we use the bad words lexicon described in Sec. 2.
	TW-LING-QST	Percentage of tweets containing at least one question (i.e. sentence with a question mark).
	TW-LING-LEV	Average Levenshtein distance between tweets.
	TW-LING-ENG	Percentage of tokens which match any word in the English dictionary described in Sec. 2.
	TW-LING-ENT-OC	Average number of mentions of the target entity across all tweets.
	TW-LING-ENT-VB	Percentage of verbs whose corresponding subject is the target entity.
	TW-LING-ENT-TW	Percentage of tweets containing at least one verb whose subject is the target entity.
Structural	TW-STRC-TOK	Number of tokens in the snapshot.
	TW-STRC-TWE	Number of tweets in the snapshot.
	TW-STRC-RET	Percentage of tweets that are retweets.
	TW-STRC-REP	Percentage of tweets that are replies.
	TW-STRC-USR	Average number of tweets per user.
	TW-STRC-TIM	Two features, representing mean and std.dev. of the distribution modeling tweets' timestamps.
	TW-STRC-HST	Number of unique hashtags with respect to the total number of hashtags.
Buzziness	TW-BUZZ	Estimates entity buziness: TW-BUZZ $= \frac{\|s\|}{(\sum_{i \in prev(s,N)} \|s_i\|)/N}$, where $\|s\|$ is the number of tweets in the snapshot; $prev(s)$ are snapshots referring to the same entity of s, in N time periods previous to s (we set $N = 2$).
Sentiment	TW-SENT-POS	Fraction of positive tweets (i.e. $pol(t) > 0$)
	TW-SENT-NEG	Fraction of negative tweets (i.e. $pol(t) < 0$)
	TW-SENT-NEU	Fraction of neutral tweets (i.e. $pol(t) = 0$)
Controversy	TW-CONT-MIX	Estimate how many mixed positive and negative tweets are in the snapshot: TW-CONT-MIX $= \frac{Min(\|Pos\|,\|Neg\|)}{Max(\|Pos)\|,\|Neg\|)} \cdot \frac{\|Pos\|+\|Neg\|}{\|Pos\|+\|Neg\|+\|Neu\|}$, where Pos, Neg and Neu are the sets of tweets with positive,negative and neutral polarity.
	TW-CONT-TSY	The contradiction score adopted by [7]: TW-CONT-TSY $= \frac{\theta \cdot \sigma^2}{\theta + (\mu)^2} \cdot W$, where μ and σ^2 are respectively the mean and the variance of the polarity scores $pol(t)$ of the tweets; parameters θ and W are set as in [7].
	TW-CONT-HST	Four features, representing the fraction over the total number of hashtags in the snapshot, of the following hashtags: '#controv', '#scandal', '#unheard' and '#wft'.
	TW-CONT-TRM	Percentage of tweets with least one controversy word from our controversy lexicon in Sec. 2.
External features		
News buzz	EX-BUZZ-1	Number of articles aligned with the given snapshot.
	EX-BUZZ-2	Change in the amount of news coverage for the given entity with respect to the recent past: EX-BUZZ-2 $= \frac{\|articles\| - (\sum_{1<i<N} \|articles_i\|)/N}{\|articles\|}$, where $\|articles_i\|$ is the number of articles about the target entity in a particular time period previous to s (we use $N = 7$).
Web-News controversy	EX-CONT-HIST	Controversy level of an entity in Web data: EX-CONT-HIST $= k/\|controversyLexicon\|$, where k is the number of terms in our controversy lexicon, whose co-occurrence pointwise mutual information with the target entity on the Web, is above 2; and $\|controversyLexicon\|$ is the size of the controversy lexicon.
	EX-CONT-ASS-1	Sum of overall controversy scores (EX-CONT-HIST) for the entities co-ocurring with the target entity in the aligned news article set.
	EX-CONT-ASS-2	Average of overall controversy scores (EX-CONT-HIST) for the entities co-ocurring with the target entity in the aligned news article set.
	EX-CONT-TRM-1	Average number of controversy terms per news article (over all articles aligned with the snapshot)
	EX-CONT-TRM-2	Max number of controversy terms per news article (over all articles aligned with the snapshot).
	EX-CONT-TRM-3	Number of articles aligned with the snapshot that contain controversy terms .

Table 2: Feature space used in our models.

non-event snapshots and 325 are event snapshots (kappa-agreement is 1.00). Of the latter, 152 are controversial-event snapshots, and 173 non-controversial-event snapshot (kappa-agreement is 0.89, corresponding to *almost perfect agreement*). The final gold standard for our task thus contains 152 positive examples (controversial-event snapshots) and 648 negative examples.

Evaluation metrics. We compare the models on the task of ranking snapshots according to their controversy score. We employ two measures commonly used to evaluate ranking quality, *average precision* (AP) and *area under the ROC curve* (AROC) [3], in a 10-fold cross-validation setup. The AROC represents the probability that a model will rank a controversial-event snapshot higher than a randomly chosen snapshot. We compare the following systems:

base-rnd. A random-rank baseline.

base-hst. A ML model using EX-CONT-HIST as the only feature.

base-tsy. A ML model using only the TW-CONT-TSY feature from Tsytsarau et al. [7].

direct. Our ML direct model using all our features.

pipeline. Our ML two-step pipeline model using all features in both steps.

blended. Our ML blended model using all features.[5]

3.1 Experimental Results

Table 4 contains the experimental results: all our models outperform the baselines with statistical significance at the 0.95 level. The **blended** system performs best, indicating

[5]For all models, GBDT parameters were set on an independent development set, as follows: number of trees=50, shrinkage= 0.01, max nodes per tree=10, sample rate=0.5

ENTITY	TIME PERIOD	EXAMPLE TWEETS	EVENT
Kelly Clarkson	23/11/2009	*"kelly clarkson has the best songs"* *"is it just me or did kelly clarkson mess up the line on the 1st verse of Already Gone ???"*	Performance at American Music Awards (*controversial event*)
Joe Biden	28/01/2010	*"joe biden a little too quick on the applause #sotu"* *"omg i love how joe biden would like to clap after every sentence and makes the best facial expressions ever"* *"i heart joe biden. rock it obama !"*	Co-presiding during State of Union (*controversial event*)

Table 3: Examples of top-ranked snapshots for the best-performing blended model.

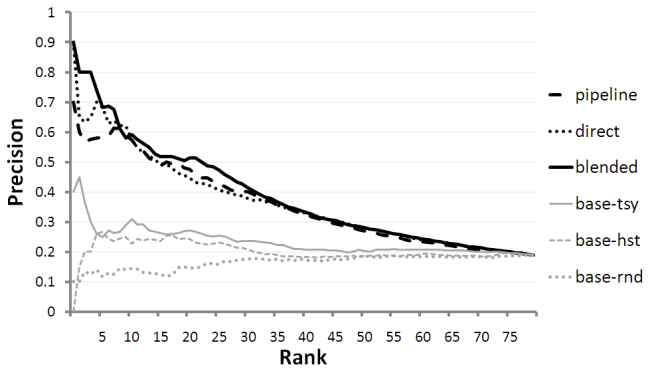

Figure 1: Precision at rank for different models, averaged over the 10 folds.

that a 'soft' integration between the 1st and 2nd steps (event detection and controversy detection), is preferred to a 'hard' integration (**pipeline** model) or a direct approach (**direct** model). The lower performance of **pipeline** is due to the discarding of too many event snapshots in the event detection step. However, the differences in performance between the systems are not statistically significant at the 0.95 level.

The overall AROC results show that our models have good discriminative power - they are able to rank controversial-event snapshots higher than non-controversial-event snapshots with probability greater than 0.80.

Figure 1 plots the precision-at-rank curves, while Table 3 reports the top-ranked snapshots for the **blended** model. The **blended** model has a very high precision at high ranks: precision at rank-1 is 0.90, and at rank-4 is 0.80. Projecting these numbers to the overall size of the snapshot set (73,368 examples), we estimate that **blended** extracts the top-900 controversial snapshots with 0.90 precision, and the top-3,700 with 0.80 precision. This is encouraging, especially in view of plugging in the models into a real application.

Feature analysis. The top-10 most discriminative features[6] for the GBDT **blended** model are a diverse mix representing various feature families. As expected, EVENT-SCORE is the most relevant feature for the blended model, as it takes care of separating event snapshots from non-event ones. TW-STRC-HST ranks second, suggesting that in general hashtags are an important semantic component of tweets: they help identify the topic of a tweet and estimate the topical cohesiveness of a set of tweets. External features based on news and the Web are also useful: coupling Twitter information with traditional media helps validate and explain social media reactions. Linguistic, structural and sentiment features are also highly ranked, which indicates that a rich, varied set of features is key for controversy detection.

[6]As ranked by GBDT [2], features with higher importance have a greater contribution in the model building phase.

Error Analysis. Our error analysis found that *false positives* (i.e. highly but incorrectly ranked non-controversial-event snapshots) are mostly snapshots containing harder-to-interpret, misleading mixtures of positive and negative words (e.g., negative words used in a positive sense, as in "my boy shannon brown is killing it".) We plan to integrate strategies for ambiguity resolution to deal with such cases.

Causes for *false negatives* (controversial-event snapshots incorrectly ranked low) include negative terms used in a positive sense (e.g. "roy williams crazy as hell lol").

Future work Our future work plans involve improved tweet-level sentiment detection, additional features for controversy detection and an automatic analysis of the identified events (e.g., personal vs. professional developments, lasting vs. short-lived controversies). We also plan the integration of our research in applications: detecting recent events about well-known entities can be useful to improve user's search experience and user's engagement on a site.

4. REFERENCES

[1] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21:543–565, 1995.

[2] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[3] D. Green and J. Swets. *Signal detection theory and psychophysics*. Wiley, 1966.

[4] G. Mishne and N. Glance. Leave a Reply: an Analysis of Weblog Comments. In *Third Annual Workshop on the Weblogging Ecosystem, WWW*, 2006.

[5] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP-2005*, pages 339–346, 2005.

[6] H. Sayyadi, M. Hurst, and A. Maykov. Event Detection and Tracking in Social Streams. In *Proceedings of ICWSM*, 2009.

[7] M. Tsytsarau, T. Palpanas, and K. Denecke. Scalable Discovery of Contradictions on the Web. In *Proceedings of WWW*, 2010.

[8] J. Wiebe and C. Cardie. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, 2005.

[9] Q. Zhao, P. Mitra, and B. Chen. Temporal and Information Flow Based Event Detection from Social Text Streams. In *Proceedings of the WWW*, 2007.