Detecting controversies in Twitter: a first study

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

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

Social media gives researchers a great opportunity to understand how the public feels and thinks about a variety of topics, from political issues to entertainment choices. While previous research has explored the likes and dislikes of audiences, we focus on a related but different task of detecting controversies involving popular entities, and understanding their causes. Intuitively, if people hotly debate an entity in a given period of time, there is a good chance of a controversy occurring. Consequently, we use Twitter data, boosted with knowledge extracted from the Web, as a starting approach: This paper introduces our task, an initial method and encouraging early results.

Controversy Detection. We focus on detecting controversies involving known entities in Twitter data. Let a snapshot denote a triple \( s = (e, \Delta t, tweets) \), where \( e \) is an entity, \( \Delta t \) is a time period and \( tweets \) is the set of tweets from the target time period which refer to the target entity.\(^1\) Let \( cont(s) \) denote the level of controversy associated with entity \( e \) in the context of the snapshot \( s \). Our task is as follows:

Task. Given an entity set \( E \) and a snapshot set \( S = \{(e, \Delta t, tweets) | e \in E\} \), compute the controversy level \( cont(s) \) for each snapshot \( s \) in \( S \) and rank \( S \) with respect to the resulting scores.

Overall Solution. Figure 1 gives an overview of our solution. We first select the set \( B \subset S \), consisting of candidate snapshots that are likely to be controversial (buzzy snapshots). Then, for each snapshot in \( B \), we compute the controversy score \( cont \), by combining a timely controversy score \( (tcont) \) and a historical controversy score \( (hcont) \).

Resources. Our method uses a sentiment lexicon \( SL \) (7590 terms) and a controversy lexicon \( CL \) (750 terms). The sentiment lexicon is composed by augmenting the set of positive and negative polarity terms in OpinionFinder 1.5 \(^2\) (e.g. ‘love’, ‘wrong’) with terms bootstrapped from a large set of user reviews. The controversy lexicon is compiled by mining controversial terms (e.g. ‘trial’, ‘apology’) from Wikipedia pages of people included in the Wikipedia controversial topic list.

Selecting buzzy snapshots. We make the simple assumption that if in a given time period, an entity is discussed more than in the recent past, then a controversy involving the entity is likely to occur in that period. We model the intuition with the score:

\[
b(s) = \frac{|tweets_s|}{\left( \sum_{i \in prev(s,N)} |tweets_i| \right)/N}
\]

where \( tweets_s \) is the set of tweets in the snapshot \( s \); and \( prev(s,N) \) is the set of snapshots referring to the same entity of \( s \), in \( N \) time periods previous to \( s \). In our experiment, we use \( N = 2 \), i.e. we focus on two days before \( s \). We retain as buzzy snapshots only those with \( b(s) > 3.0 \).

Historical controversy score. The \( hcont \) score estimates the overall controversy level of an entity in Web data, independently of time. We consider \( hcont \) our baseline system, to which we compare the Twitter-based models. The score is estimated on Web document data using the \( CL \) lexicon as fol-

---

\(^1\)We use 1-day as the time period \( \Delta t \). E.g. \( s = ('Brad Pitt',12/11/2009,tweets) \)

lows: $hcont(e) = k/|CL|$, where $k$ is the number of controversy terms $t'$ s.t. $PMI(e, t') > A^3$.

**Timely controversy score.** $tcont$ estimates the controversy of an entity by analyzing the discussion among Twitter’s users in a given time period, i.e. in a given snapshot. It is a linear combination (tuned with $\alpha \in [0, 1]$) of two scores: $MixSent(s)$: reflects the relative disagreement about the entity in the Twitter data from snapshot $s$. First, each of the $N$ tweets in $s$ is placed in one of the following sets: Positive (Pos), Negative (Neg), Neutral (Neu), based on the number of positive and negative SL terms in the tweet. $MixSent$ is computed as:

$$MixSent(s) = \frac{\min(|\text{Pos}|, |\text{Neg}|)}{\max(|\text{Pos}|, |\text{Neg}|)} \times \frac{|\text{Pos}| + |\text{Neg}|}{N}$$

$Controv(s)$: this score reflects the presence of explicit controversy terms in tweets. It is computed as: $Controv(s) = |ctv|/N$, where $ctv$ is the set of tweets in $s$ which contain at least one controversy term from $CL$.

**Overall controversy score.** The overall score is a linear combination of the timely and historical scores: $cont(s) = \beta \cdot tcont(s) + (1 - \beta) \cdot hcont(s)$, where $\beta \in [0, 1]$ is a parameter.

**Experimental Results**

We evaluate our model on the task of ranking snapshots according to their controversy level. Our corpus is a large set of Twitter data from Jul-2009 to Feb-2010. The set of entities $E$ is composed of 104,713 celebrity names scraped from Wikipedia for the Actor, Athlete, Politician and Musician categories. The overall size of $S$ amounts to 661,226 (we consider only snapshots with a minimum of 10 tweets). The number of buzzy snapshots in $B$ is 30,451. For evaluation, we use a gold standard of 120 snapshots randomly sampled from $B$, and manually annotated as controversial or not-contraversial by two expert annotators (detailed guidelines will be presented at the workshop). Kappa-agreement between the annotators, estimated on a subset of 20 snapshots, is 0.89 (‘almost perfect’ agreement). We experiment with different $\alpha$ and $\beta$ values, as reported in Table 1, in order to discern the value of final score components. We use Average Precision (AP), and the area under the ROC curve (AROC) as our evaluation measures.

The results in Table 1 show that all Twitter-based models perform better than the Web-based baseline. The most effective basic model is $MixSent$, suggesting that the presence of mixed polarity sentiment terms in a snapshot is a good indicator of controversy. For example, ‘Claudia Jordan’ appears in a snapshot with a mix of positive and negative terms -in a debate about a red carpet appearance- but the $hcont$ and $Controv$ scores are low as there is no record of historical controversy or explicit controversy terms in the target tweets. Best overall performance is achieved by a mixed model combining the $hcont$ and the $MixSent$ score (last row in Table 1). There are indeed cases in which the evidence from $MixSent$ is not enough - e.g., a snapshot discussing ‘Jesse Jackson’’s appearance on a tv show lacks common positive or negative terms, but reflects users’ confusion nevertheless; however, ‘Jesse Jackson’ has a high historical controversy score, which leads our combined model to correctly assign a high controversy score to the snapshot. Interestingly, most controversies in the gold standard refer to micro-events (e.g., tv show, award show or athletic event appearances), rather than more traditional controversial events found in news streams (e.g., speeches about climate change, controversial movie releases, etc.); this further strengthens the case that Twitter is a complementary information source wrt news corpora.

We plan to follow up on this very preliminary investigation by improving our Twitter-based sentiment detection, incorporating blog and news data and generalizing our controversy model (e.g., discovering the ‘what’ and the ‘why’ of a controversy, and tracking common controversial behaviors of entities over time).