Sentiment Analysis on Evolving Social Streams: How Self-Report Imbalances Can Help

Pedro Calais Guerra, Wagner Meira Jr.
Dept. of Computer Science
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG, Brazil
{pcalais,meira}@dcc.ufmg.br

Claire Cardie
Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
cardie@cs.cornell.edu

ABSTRACT

Real-time sentiment analysis is a challenging machine learning task, due to scarcity of labeled data and sudden changes in sentiment caused by real-world events that need to be instantly interpreted. In this paper we propose solutions to acquire labels and cope with concept drift in this setting, by using findings from social psychology on how humans prefer to disclose some types of emotions. In particular, we use findings that humans are more motivated to report **positive** feelings rather than **negative** feelings and also prefer to report **extreme** feelings rather than **average** feelings.

We map each of these *self-report imbalances* on two machine learning sub-tasks. The preference on the disclosure of positive feelings can be explored to generate labeled data on *polarizing* topics, where a positive event for one group usually induces negative feelings from the opposing group, generating an imbalance on user activity that unveils the current dominant sentiment.

Based on the knowledge that extreme experiences are more reported than average experiences, we propose a feature representation strategy that focus on terms which appear at spikes in the social stream. When comparing to a static text representation (TF-IDF), we found that our feature representation is more capable of detecting new informative features that capture the sudden changes on sentiment stream caused by real-world events.

We show that our social psychology-inspired framework produces accuracies up to 84% while analyzing live reactions in the debate of two popular sports on Twitter – soccer and football – despite requiring no human effort in generating supervisory labels.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Experimentation

Keywords

Sentiment Analysis, Stream Data Mining, Social Media Analytics

WSDM'14, February 24–28, 2014, New York, New York, USA.
Copyright 2014 ACM 978-1-4503-2351-2/14/02 ...$15.00.
http://dx.doi.org/10.1145/2556195.2556261.

1. INTRODUCTION

One goal of a sentiment analysis system is to, given a text document, infer its polarity toward entities and events mentioned in the text [45, 32]. As social media platforms become the primary medium used by people to express their opinions and feelings about a multitude of topics that pop up daily on news media, the vast amount of opinionated data now available in the form of *social streams* gives us an unprecedented opportunity to build valuable applications that monitor public opinions and opinion shifts [24, 23]. For example, a sports web portal can track the crowd sentiment during live matches, something far more appealing than the relative number of mentions of each team, which is what most sports web sites currently offer. Creating such applications enrich the personal experience of watching live events on TV, and following the social media buzz simultaneously with live broadcasted events is becoming a joint experience, where watching not only the event itself, but how others react to it, is part of the experience.

The task of interpreting positive and negative feelings expressed on social streams exhibits a number of unique characteristics that are not present in the static and well-controlled domains on which sentiment analysis has focused in the last decade – mainly product and movie reviews [45, 37, 23]. On the downside, it faces two challenges that are common to many data stream classification tasks [33]: (i) the limited availability of labeled data and (ii) the need to deal with the evolving nature of the stream, which causes the target concept to change and requires learning models to be constantly updated – a problem known as *concept drift* [47]. Challenge (i) is a serious drawback because current sentiment analysis models are heavily based on supervised approaches [37, 45], and human constraints on generating a constant flow of labeled messages on streams remain high. The sparsity of language, the use of neologisms and word lengthening as an indicator of sentiment (e.g., "cooooooooool!", "goooooooooaal!" [8]) also contribute to make the process of acquiring large labeled sets of pre-classified messages unfeasible [23]. Challenge (ii) arises in sentiment streams as it is necessary to deal with constant changes of vocabulary and sudden changes of sentiment in reaction to real-world events. For example, in a few minutes a positive sentiment of the fans of a soccer team commenting on Twitter or Facebook may vanish by a goal scored by the adversary team; such *sentiment drift* represents a great challenge for real-time sentiment tracking, since it requires the stream classifier to be capable of quickly identifying and adapting to the sudden change on the dominant sentiment [43].

Despite these important constraints and drawbacks, streams reflecting the society's immediate emotional reactions regarding a topic have an important property, which we seek to exploit in this work, namely, the flow of opinions from social networking services is inherently constrained to manifestations from individuals that have explicitly and deliberately *chosen* to post a message in reaction to some real-world event; thus, the distribution of positive and negative opinions is potentially quite different from the random samples obtained in traditional opinion polls and survey methodologies [31]. Although such *reporting bias* is usually perceived as a source of inaccuracy [28, 16], here we argue that the self-reporting nature of social media, when observed on large-scale social network data, may actually provide signals that ease the task of sentiment tracking in online environments, provided that we understand the **factors** that motivate people to publicly express their feelings. We build sentiment analysis models that exploit two factors widely described by substantive research from social psychology and behavioral economics that describe human preferences when disclosing emotion publicly:

> **Positive-negative sentiment report imbalance:** People tend to express positive feelings more than negative feelings in social environments [5, 12, 30, 25].

> **Extreme-average sentiment report imbalance:** People tend to express extreme feelings more than average feelings in social environments [2, 11, 10, 28].

We explore each of these two self-report imbalances to accomplish a different subtask in learning-based sentiment analysis. The first self-report factor, which we call **positive-negative sentiment report imbalance** throughout the paper, is employed to acquire labeled data that supports supervised classifiers. In the context of *polarizing groups* – a division of the population into groups of people sharing similar opinions in the context of a topic [3, 19], a positive event for one group tends to be negative to the other, and vice-versa. For example, while supporters of a football team are likely to be happy when their team scores, fans of the adversary team are expected to be upset when faced with the same event. Based on social psychology research that states that the disclosure of positive feelings is preferred, we can then make a prediction of the current dominant sentiment by simply counting how many members of each group, relative to group sizes, decided to post a message during the specified time frame. Since the social context information only holds during time frames when a significant real-world event happens, we adopt a probabilistic model that computes the uncertainty of the social context, and, at each time frame, generates a probabilistic sentiment label, which can then be incorporated into a range of content-based supervised classifiers.

The second self-report factor we explore is related to the human tendency to report extreme experiences more than average experiences [2, 11, 10, 28]. The **extreme-average sentiment report imbalance** implies an important consequence for real-time sentiment tracking: because extreme feelings stimulate reactions, spikes of activity in streams of opinionated text tend to contain highly emotional terms, which are precisely the features that are helpful for sentiment prediction. We propose a simple text representation strategy based on this observation, named *term arousal*, that maintains, for each term (or lexical unity, e.g., n-grams), a measure of how often it appears in high-volume time windows in the stream; we call these **high-arousal** terms. Our experimental studies demonstrate that these terms are better indicators of emerging and strong feelings than traditional static representations (e.g., TF-IDF), allowing the underlying classification model to adapt quicker to sudden sentiment drift induced by real-world events.

In summary, our main contributions in this paper are:

1. We raise awareness over the fact that opinions expressed on social media platforms are *not* a random sample of the online population, but are impacted by many social and psychological factors that need to be accounted for in order to build reliable and useful sentiment analysis systems;

2. We show that self-report imbalances create rich *social contexts* that can be leveraged to improve two key subtasks in the construction of a sentiment stream classifier – namely, the acquisition of labeled data and feature representation suitable to deal with sudden sentiment drifts.

We evaluated our social psychology-inspired framework on sports events heavily debated on Twitter; when instantiating our framework with a Multinomial Naive Bayes classifier, our results are comparable to what is typically obtained as an acceptable result for document-level sentiment analysis – between 80% and 85% of accuracy [45] – but, because the stream-based scenario imposes stricter and harder constraints, we believe they point to a promising option for sentiment classification on evolving social streams. In addition, our approach targets two generic sub-tasks for learning-based sentiment analysis – label acquisition and feature representation. As a result, our framework can be incorporated into sophisticated sentiment classifiers that make use of more powerful NLP models and features.

2. SOCIAL PSYCHOLOGY BACKGROUND

Psychologists classify emotions into two independent dimensions: pleasure (happiness or sadness) and activation (or arousal) [4]. The *self-report imbalances* we briefly presented in Section 1 are biases in the bidimensional emotion space caused by the fact that social media systems are *communicative* platforms; as a consequence, opinions and feelings expressed in online social environments are a result of opinion holders' explicit desire to make his friends or followers aware of his or her opinions. In other words, the communicative nature of social media makes social data a side effect of intentional and deliberate communication between users, rather than as a representation of some underlying activity [39, 31].

On the positive-negative dimension, the preference on the disclosure of positive feelings is caused by our need in being perceived as successful and happy persons [34, 40], and it causes a bias where everyone in online social environments perceives others as happier than they actually are [25]. In the case of opinions expressed over a polarizing topic, the preference on sharing positive news and opinions goes beyond the human's desire to improve his or her reputation: each group also gives preference to news and facts that favor their viewpoints, a result of many biases such as *confirmation bias* and *selective exposure* [20, 31]. Notice that the definition of a *positive* event is group-dependent: for rival supporters of a team or opposers of politicians in office, negative facts such as a conceded goal or a political scandal will be explored by them as "positive" – i.e., as a motivation to explore the fact to their benefit. Also, in some contexts, such as product reviews, the bias leans toward the disclosure of negative experiences [22]; our sentiment analysis framework is generalized to take advantage of the asymmetry on either direction.

On the arousal dimension, it was found that extreme emotions – angriness, anxiety, awe, excitement – are *high-arousal* emotions: they affect our body and put us in a state of activation and readiness for action [4, 5]. In social media, action means making private feelings public, what makes sentiment expressed on online media to be biased towards strong feelings and opinions.

In the next sections we will detail how we embed these biases on sentiment self-report in the analysis of feelings expressed on social streams.

3. ACQUIRING LABELED DATA

Differently from the majority of research on supervised sentiment analysis, which focus on batch processing of opinionated documents [37, 45], here we are interested in the setting where the data arrives as an infinite stream and reflects real-world unpredictable events. As we discussed in Section 1, in this setting a constant flow of labeled messages is required to build and update supervised sentiment models. Unfortunately, in textual streams characterized by sparse and time-changing content it is not feasible to manually obtain labeled data in significant amounts and in a timely manner [33].

To overcome this problem, we propose a method to acquire labeled messages by exploring the *positive-negative sentiment report imbalance* in the context of polarizing groups. We compare the strength of reactions of polarizing groups during each time span, moving from processing individual messages to processing groups of messages. These groups are obtained by dividing the social stream into a sequence of non-overlapping and contiguous time windows of equal duration (e.g., Δt minutes), what gives us the capability of exploiting the **social context** induced by the set of users that expressed their sentiment w.r.t. topic T during each time window W_t. Each window W_t contains all messages sent during the time period $[t_i, t_i + \Delta t]$ (W_0 starts at t_0 and $t_{i+1} = t_i + \Delta t$) and is composed of a triple $(\mathcal{S}_t, \mathcal{D}_t, \mathcal{Y}_t)$:

- $\mathcal{S}_t$ is a multiset of group memberships of all users who posted a message during W_t. On a polarized domain, we assume that each user belongs to one of two groups, G_A or G_B.[1] For instance, $\mathcal{S}_t = \{G_A, G_A, G_A, G_B, G_B\}$ indicates that 3 members of group G_A and 2 members of group G_B posted a message during W_t. Assigning users to groups is a task that can be accomplished by several community detection and graph mining techniques that explore the social ties among users, under the assumption that similar users are likely to connect to each other [1, 19].

- $\mathcal{D}_t$ is the sum vector of all feature vectors extracted from messages written during W_t;

- $\mathcal{Y}_{e,t} \in \{+, -\}$ indicates the ground-truth sentiment expressed during W_t w.r.t. an entity e in the context of topic T. Here, each e is an individual or organization naturally linked to the polarizing group that supports it; for instance, if $G = \{\text{Democrats}\}$, than $e(G) = \{\text{Barack Obama}\}$, and $e(G) = \{\text{New York Giants team}\}$ if $G = \{\text{New York Giants fans}\}$.

Note that, instead of seeking for labels for individual messages, we label *all* the messages mentioning an entity e in time window W_t with the same polarity $\mathcal{Y}_{e,t}$. Although we do not expect every opinion expressed during a time window to follow the same polarity, we seek here to determine the *dominant* sentiment during W_t; furthermore, the probabilistic method we will detail next assigns a confidence on the label estimation, what can be interpreted as an estimate of the proportion of positive and negative messages written during a given W_t.

For now we ignore the content vector $\mathcal{D}_t$ and focus on $\mathcal{S}_t$ as an input to build a sentiment prediction function $f : \mathcal{S} \to \mathcal{Y}$. The fundamental principle we seek to exploit is that, on polarized discussions dominated by two opposing groups G_A and G_B, in general $\mathcal{Y}_{e(G_A),t} = +$ implies that $\mathcal{Y}_{e(G_B),t} = -$, and vice-versa (we will relax this requirement in Section 4, by learning a content-based classifier based on labels provided by $\mathcal{S}_t$). A simple approach to predict $\mathcal{Y}_t$ based on $\mathcal{S}_t$ is to consider that each message is a "vote" toward the sentiment expected to drive more reactions and, thus, a majority-voting strategy is employed to predict the dominant sentiment at W_t. In the toy example $\mathcal{S}_t = \{G_A, G_A, G_A, G_B, G_B\}$, since we are supported by social theories that indicate preference toward the report of positive sentiment, we would predict 3 votes for labels ($\mathcal{Y}_{e(G_A),t} = +$, $\mathcal{Y}_{e(G_B),t} = -$) and 2 votes for labels ($\mathcal{Y}_{e(G_A),t} = -$, $\mathcal{Y}_{e(G_B),t} = +$). The only point of caution here is that normalizing by group sizes $|G_A|$ and $|G_B|$ is important to discount the effect of larger groups on $\mathcal{S}_t$.

Majority-voting is a simple and straight-forward approach, but it has an important limitation: it does not quantify the uncertainty on the information provided by the voters [42]. Since the labeling mechanism by social context is not perfect, capturing the degree of confidence on the correlation between $\mathcal{S}_t$ and $\mathcal{Y}_t$ is crucial if we will incorporate this information on learning models. In particular, the labeling scheme based on positive-negative report imbalance is error-prone due to two reasons:

1. $\mathcal{S}_t$ is likely to carry a significant correlation with the dominant sentiment only when a well-determined and relevant event happened during time window W_t, i.e., a goal or touchdown in a sports match, or some breaking news on the topic being followed. Most of the time, the positive-negative report imbalance will not be triggered at a sufficient strength, and an unreliable prediction will be generated.

2. Since we are modeling only user posting decisions in face of positive/negative events and abstracting from several other factors that influence the posting decision (as well as different individual posting probabilities), we are prone to deal with noise due to the many factors that motivate user reactions and that we are not accounting for.

Therefore, in order to make our approach reliable and more useful, it is desirable to associate with each predicted label $\mathcal{Y}_t$ a measure of confidence $P(\mathcal{Y}_t|\mathcal{S}_t)$ that captures the noisy nature of the multiset of group memberships $\mathcal{S}_t$. We instantiate a probabilistic model that assumes that on each time window W_t a coin of bias θ_t is tossed to decide whether each message will be authored by a member of G_A or G_B, and $|G_{A,t}|$ messages from members of G_A and $|G_{B,t}|$ from members of G_B are observed. A fair coin is expected to generate a number of heads (G_A) and tails (G_B) proportional to $\theta_{fair} = \frac{|G_A|}{|G_A|+|G_B|}$ and $1 - \theta_{fair}$, respectively, modeling the fact that members of both groups are reporting their sentiment with the same probability. Alternatively, a biased coin, whose θ_t is different from $\frac{|G_A|}{|G_A|+|G_B|}$ at some degree, means that members of one group are self-reporting their feelings at a higher rate than the other, indicating that its members are probably experiencing positive feelings in comparison to the other group.

A coin model is convenient because it naturally models the intuitive fact that spikes of activity in the social stream are more informative: in the same way that our confidence on the bias of a coin increases as we toss it more times, a time window W_t which contains a large number of messages (and, consequently, a larger multiset $\mathcal{S}_t$) is more likely to carry a clear dominant sentiment, not only due to a larger sample, but because spikes of activity are likely to be associated with real-world events that trigger the positive-negative report imbalance. Our probabilistic model is divided into two steps:

1. Estimate the probability distribution on the latent variable θ_t;

2. Estimate how far θ_t is from $\theta_{fair} = \frac{|G_A|}{|G_A|+|G_B|}$.

[1] In practice, a domain can be associated with more than two groups, i.e., N=20 groups of supporters are found on National Football League. However, at each event of interest (e.g., a football match), we focus on the two polarizing groups that have a direct interest on it.

We use Bayesian estimates in both steps. To estimate the uncertainty on θ_t, we need to calculate the posterior predictive distribution $P(\theta_t|\mathcal{S}_t)$, i.e., the distribution over θ_t after observing the resulting multiset $\mathcal{S}_t$. In Bayesian inference, the posterior $P(\theta_t|\mathcal{S}_t)$ is proportional to a likelihood function $P(\mathcal{S}_t|\theta_t)$ and a prior distribution $P(\theta_t)$; we adopt the classical Beta-Binomial model: $P(\mathcal{S}_t|\theta_t)$ is computed from a binomial distribution $Bin(|W_t|, \frac{|G_{A,t}|}{|G_{A,t}|+|G_{B,t}|})$ and the prior follows a Beta distribution $Beta(a, b)$ (a and b are hyperparameters) [42, 7]. As a result of the conjugacy property of the Binomial and the Beta distributions, the posterior predictive distribution nicely follows a Beta distribution $Beta(|G_{A,t}|+a, |G_{B,t}|+b)$ that captures our uncertainty over θ_t [7].

It is still necessary to choose the hyperparameters a and b that govern the prior distribution $P(\theta_t)$ and capture the knowledge acquired from previous observed data streams over the noisy nature of the coin. To incorporate our prior knowledge that θ_t is expected to be proportional to group sizes, we want to find hyperparameters a and b in the form $a = \frac{K|G_A|}{|G_A|+|G_B|}$ and $b = \frac{K|G_B|}{|G_A|+|G_B|}$. K can be understood as a smoothing parameter: the greater its value, the more confident the model is that θ_t is close to θ_{fair} and less importance will be given to the data. On the other hand, if we choose an uniform prior $Beta(1, 1)$, then we let the model rely totally on the observed data to judge how likely the tosses are coming from a coin of bias θ_t; the expected value of the coin bias in this case is equivalent to the maximum likelihood estimate $\theta_t = \frac{|G_{A,t}|}{|G_{A,t}|+|G_{B,t}|}$ [7]. Such direct estimation of θ_t makes the unrealistic assumptions that tosses are generated i.i.d. from a noiseless coin.

We estimate K from the streaming data by employing an Empirical Bayes approach[2]. To learn the extent to which the coin we are modeling is noisy, we take advantage of the data continuity in the stream: we observe a sequence of noisy estimates $(\theta_0, \theta_1, ..., \theta_t)$ of a different coin being tossed at each time window. The property we want to explore here is that we expect consecutive time windows W_i and W_{i+1} of *similar message volume* to share a similar θ; large differences in θ between these windows should be attributed to noise, since no significant real-world event has happened (otherwise we would observe a large $||S_{i+1}| - |S_i||$). On the other hand, we would like to allow consecutive time windows with a large difference in message volume to exhibit a larger absolute difference $|\theta_{i+1} - \theta_i|$, since, according to our user behavior model, a spike of activity will trigger a bias either on G_A or G_B.

We seek to find the value of K that maximizes Equation 1. ρ is the Pearson correlation coefficient, and ΔV and $\Delta\theta(K)$ are vectors containing the sequence of $||S_{i+1}| - |S_i||$ and $|\theta_{i+1} - \theta_i|$ observed on the stream. Note that we write $\Delta\theta(K)$ as a function of K, since the estimates of θ_t are affected by the prior distribution $P(\theta_t|K)$. The highest Pearson correlation will explain larger differences in θ through larger differences in time-window volume, and we estimate it by using a standard gradient descent method.

$$K = argmax(\rho(\Delta V, \Delta\theta(K))) \quad (1)$$

Recall that our goal is to estimate how far the latent variable θ_t is from $\theta_{fair} = \frac{|G_A|}{|G_A|+|G_B|}$, what indicates a bias in the posting decision of either G_A or G_B. This value can be estimated by calculating the area under the curve of the distribution $Beta(|G_{A,t}| + a, |G_{B,t}| + b)$ at the decision threshold $x = \frac{|G_A|}{|G_A|+|G_B|}$. If $I_x(a, b)$ is the CDF of $Beta(a, b)$ in the interval $(0, x)$, then

[2]Empirical Bayes methods are approaches that estimate the prior distribution over a random variable from the data itself, rather than defining the distribution before observing any data, as on standard Bayesian inference [17].

$$conf(\theta_{fair}, \mathcal{S}_t) = \max(I_{|G_A|/(|G_A|+|G_B|)}(|G_{A,t}| + a, |G_{B,t}| + b), \\ 1 - I_{|G_A|/(|G_A|+|G_B|)}(|G_{A,t}| + a, |G_{B,t}| + b)) \quad (2)$$

where I is the regularized incomplete Beta function and can be used to compute to the cumulative distribution function in a Beta distribution [42]. The value $1 - conf(\theta_{fair}, \mathcal{S}_t)$ gives us an estimate of how likely the predicted label is trustable given the observed social context $\mathcal{S}_t$, i.e., $P(\mathcal{Y}_t|\mathcal{S}_t)$.

3.1 Experimental Evaluation using Twitter data

We evaluate the predictive power of social contexts induced by the positive-negative report imbalance on the analysis of the reactions expressed on Twitter by fans of two popular sports that generate passionate debate on social media: soccer and (American) football. Sports competitions are among the topics that generate the largest fractions of audience both in broadcasting media [46] and social media [29]; however, most initiatives taken by content portals to turn the live game experience into an online social experience are still restricted to simple tools such as the display of the most popular tweets or plots on the variation of the relative number of mentions of the playing teams. Measuring the crowd sentiment during live matches is something far more appealing and may answer relevant questions such as "do the supporters still believe in a win, despite losing the match so far?".

Table 1 gives an overview of two datasets we obtained from the Twitter data collection API. The datasets comprise fans' debate on Brazilian Soccer League seasons (2010, 2011 and 2012) and NFL (2010/11, 2011/12 and 2012/13 seasons). We chose team names and specific words of each competition as keywords. More than 35.8 million tweets from 5.6 million users have been collected in the `SOCCER` dataset, and 23 million tweets from 4.2 million users in the case of the `NFL` dataset. While tweets on Brazilian soccer are mostly in Portuguese, NFL debate is dominated by English, what gives us the possibility to experiment our model in two languages, after we build a content-based stream classifier in Section 4.

Table 1: General overview of the datasets collected from Twitter.

	Soccer	**NFL**
seasons	10-11-12	10/11, 11/12, 12/13
language	Portuguese	English
# of user groups (teams)	12	20
# of tweets	35,834,453	23,094,280
# of users	5,638,906	4,230,731
# of users w/ 1+ post/week	35,121	58,981

Before performing any sentiment prediction, we need to segment the user base into polarizing groups. In the sports domain, the natural criterion for dividing users into polarizing groups is to reflect their team preference. Several community detection and graph mining approaches that leverage social ties and social interactions can be used to accomplish this task; we manually labeled a set of users with their team preference and then used the similarities in their retweet pattern to estimate the class of unlabeled users [20].

Due to the highly-dynamic nature of sporting events, we analyze sentiment and social contexts in 1-minute time windows; larger time frames may be suitable for less dynamic domains. To generate ground-truth sentiment labels, we examined the match facts and the evolving sentiments for a number of matches in the `SOCCER` and `NFL` dataset. In addition to the match score, we manually examined the content of tweets and also included cases where the match score did not reflect the sentiment, as soccer matches that ended as null ties (0–0), but the result was enough to grant one of the teams the championship title. Although each time window is associated with a set of messages, we aim to determine the overall, global sentiment which dominates each time window, instead of individually trying to predict the polarity associated with each post.

Figure 1 shows the accuracy on the sentiment prediction task for the two datasets. On the x axis, we grouped time windows according to its volume in relation to the average time window volume: $bin = i$ corresponds to time windows where the number of messages were between i and $i + 1$ times the average.

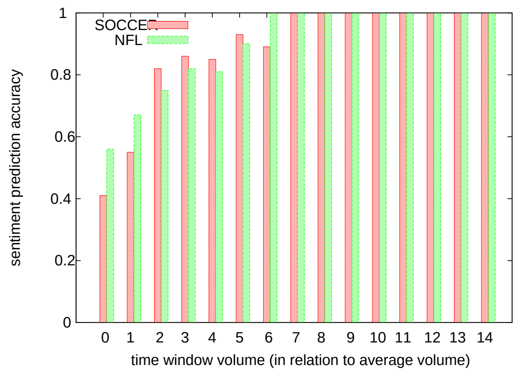

Figure 1: Accuracy on sentiment prediction on 1-minute time windows. We grouped time windows according to its volume in relation to the average time window volume. Social contexts based on positive-negative sentiment report imbalance are highly effective on sentiment prediction on large-volume time windows.

We observe that, for high-volume time windows, accuracy is very high: we could predict with more than 90% of accuracy the dominant sentiment on time windows whose volume of tweets were at least 5 times the average, despite not taking any textual content into account. This result validates the sociopsychological principle that motivated our method – positive and negative feelings are disclosed with different probabilities – and, confirms that, in the sports domain, sentiment report is biased toward the positive feeling.

We can also note from the histogram that accuracy decreases with the volume of tweets in the time-window; on time-windows whose volume is above average, accuracy is comparable to a random guesser, meaning that the induced social context is not relevant and the positive-negative report imbalance is not triggered in sufficient strength, and other factors are affecting the posting decisions' of members of G_A and G_B.

Since the majority of the time windows are not voluminous, it is important to capture the uncertainty on the sentiment prediction made by social contexts. In order to instantiate the probabilistic measure of label uncertainty we presented in this section, we use the data to set hyperparameters K_{soccer} and K_{NFL} that capture the previous knowledge on the coin that control the relationship between messages and author's groups over time. We found $K_{soccer} = 12000$ and $K_{NFL} = 6000$ as the value that maximizes the Pearson correlation that relates ΔV and $\Delta\theta(K)$ (Equation 1). Figure 2 compares, for the SOCCER dataset, the theoretical label uncertainty prediction with the empirical accuracy obtained for each volume bin; the approximation is reasonable, and results are similar for the NFL dataset.

Figure 3 shows the convex shape of the Pearson correlation measure (Equation 1) as we increase the hyperparameter K_{Soccer} in the coin model. On the red curve, we plot the absolute error between the predicted and empirical accuracy for each value of K_{soccer}, to show that the maximum of the Pearson correlations coincides with the minimum of the absolute error curve. Results are similar for the NFL dataset, and demonstrate that exploring the sequence of time-windows to smooth the measure of the coin bias θ is a simple and effective strategy.

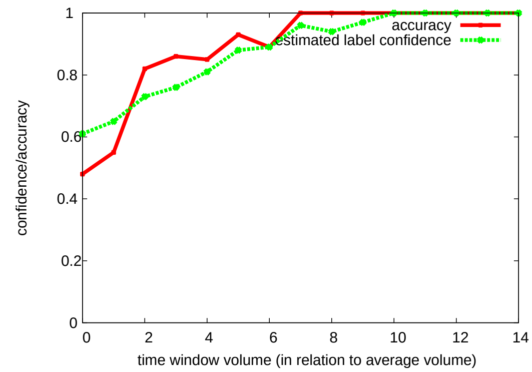

Figure 2: Difference between theoretical confidence estimate and empirical accuracy obtained for time windows of tweet volume = x times the average.

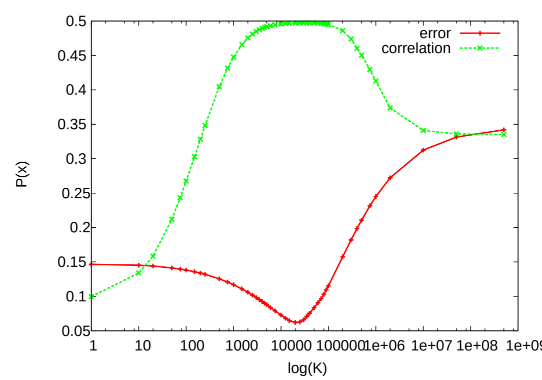

Figure 3: Choice of hyperparameter K_{soccer} as the value that maximizes Equation 1; Pearson correlation maximum coincides with the best empirical measurement of uncertainty.

4. FEATURE REPRESENTATION

In the last section, we demonstrated the predictive power of social contexts induced by the positive-negative report imbalance and the segmentation of users into polarizing groups. In addition to the low accuracy on low-time volume windows, using just $\mathcal{S}$ and ignoring content $\mathcal{D}$ is restrictive due to two reasons:

1. Sentiment prediction does not improve over time, since knowledge from past time windows is not carried to recent time windows. Improving performance as more data is processed is a basic requirement for any machine learning approach;

2. It enforces that $\mathcal{Y}_{G_A,t} = + \rightarrow \mathcal{Y}_{G_B,t} = -$, what is generally acceptable, given the polarized nature of polarized debate, but is not capable of capturing more complex variations of sentiment, where members of $|G_A|$ and $|G_B|$ can share a similar sentiment at the same time, or different intensities of sentiment.

We take inspiration on the social psychology finding that describes how humans' decision on expressing their feelings is increased by the strength of the sentiment they are experiencing [2, 11, 10, 28] (which we call, for short, as **extreme-average report imbalance**) to devise a textual feature representation (and, hence, a feature selection strategy) specially designed to track sudden variations of sentiment on evolving and dynamic social streams and that makes use of the textual feature vector $\mathcal{D}_t$ to improve accuracy on sentiment prediction.

It is widely known that the underlying text representation impacts the performance of text mining and linguistics applications [21, 44]; different *feature definition* choices (part-of-speech features, bag-of-words, n-grams etc), *feature weighting* schemes (such as binary, TF and TF-IDF) and *feature selection* approaches can be suitable for different tasks – such as text classification, text clustering and search [44, 48]. When the textual data arrives as a stream, an adequate choice of text representation is even more critical:

- The potentially infinite size of the stream limits the storage of an ever growing high dimensional feature space, what increases the need for adequate feature representation/selection that keeps the feature space as compact as possible [26].
- Static text representations (such as TF-IDF) may not be optimized to nonstationary text streams, since they do not capture adequately the dynamic nature of the feature probability distribution [27, 21], which is strongly affected by emerging new topics and real-world events.

As explained in Section 3, $\mathcal{D}_t$ is the feature vector extracted from messages written during time window W_t:

$$\mathcal{D}_t = [w_{t1}, w_{t2}, ..., w_{tM}]$$

and w_{tj} is the weight of the j-th feature in $\mathcal{D}_t$. Instead of adopting traditional term frequency (TF) or term-frequency-inverse document frequency (TF-IDF) as weights, we exploit the fact that time-windows have a varying volume of messages and, according to the extreme-average report imbalance, more people post a message when affected by an emotional, strong feeling. As a consequence, emotional content is likely to be concentrated on spikes of activity in social streams at a greater frequency than low-emotional terms. Let $\overline{W_t} = \frac{\sum_{k=0}^{t} ||W_t||}{N}$ be the average volume of messages sent in each time window up to the t-th time window and $\overline{W_{t,term}} = \frac{\sum_{k=0}^{t} ||W_k|term \in D_k||}{N_k}$ be the same measure, but considering only time windows that contain $term$. We then define $w_{t,term}$ as:

$$w_{t,term} = \frac{\overline{W_{t,term}}}{\overline{W_t}} \tag{3}$$

$w_{t,term}$ measures how the occurrence of $term$ between $[W_0, W_t]$ is correlated to high-volume time windows. $w_{t,term} = 1$ means that $term$ appears on time windows whose volume are, on average, equal to the average time window volume, and thus it indicates that the term is not expected to be associated with strong emotions (e.g., spikes). A term with $w_{t,term} = 5$ means that $term$, on average, appears on time windows whose volume are five times greater than the average. We name these terms as *high-arousal* terms, since they are associated with moments where the crowd being monitored felt motivated to react and express feelings and opinions, caused by the fact that highly emotional feelings *activate* people and drive them to action [4]).

Figure 4 provides empirical evidence that the arousal feature space is adequate to capture sentimental n-grams by correlating the arousal measure with two features commonly associated with sentiment – the use of word lengthening [8] (as on "coooooooooooool") and the use of uppercase. The more arousal we associate with a term (n-gram), the greater is the chance it is written using one of these two linguistic indicators. In Tables 2 and 3, we display the top features in each dataset according to arousal and TF-IDF. In brackets, we show the value of arousal identified for each term; high-arousal n-grams are clearly more sentimental than TF-IDF.

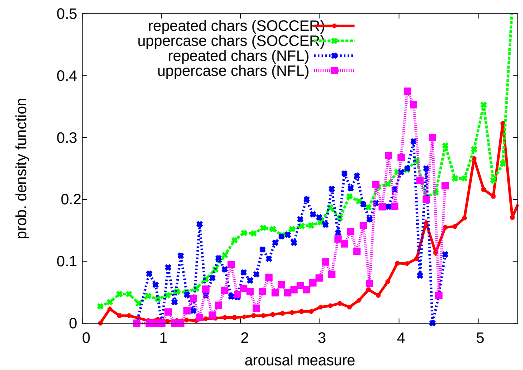

Figure 4: Indicators of excitement (use of uppercases and repeated characters) correlate with term arousal measure.

Table 2: Top 5 features for `NFL` dataset, according to *arousal* and TF-IDF representations. Arousal values are in brackets.

arousal	**TF-IDF**
PACKERS_WIN_SUPERBOWL (3.54)	yu_know_what
SUPER_BOWL_CHAMPIONS!!! (3.53)	you_would_think
YEAH! (3.43)	your_quarterback_is
superbowl_xlv_champions (2.65)	you_lost_money
touchdown!! (2.34)	you_imagine_how

Table 3: Top 5 features for `SOCCER` dataset, according to *arousal* and TF-IDF representations. Arousal values are in brackets.

arousal	**TF-IDF**
great_goal (7.53)	win!
gooooooooooooool (6.80)	gol_from_team
he_scores(5.31)	an_equalizer
GOOOL (5.00)	go!
penalty_for_team (3.34)	he_shoots

High-arousal terms and concept drift. There has been significant efforts to perform effective classification on text streams under concept drift environments; the most common strategy is to employ forgetting and weighting mechanisms that decrease the importance of old instances of data and force the stream classifier to focus on recent instances [49]. We follow a different strategy: instead of trying to restrict learning to recent examples, we design a dynamic feature space, where at any given time the feature space is defined by the useful terms selected using $arousal$ as a selection criterium. As a consequence, we are capable of quickly identifying, on spikes of activity, new features with high predictive power that may appear or gain importance over time (i.e., high values of *arousal*) that become important for sentiment classification.

When a spike occurs and (potentially) changes the dominant sentiment in the stream, due to a real world event which immediately affect users' happiness, adapting the model to such concept drift is challenging if the stream model is strongly built on past data [27]. Tackling concept drift at the feature representation stage has the advantage that unlike instance weighting and forgetting mechanisms, useful knowledge from the past is never discarded, what could harm classification perfomance [27]. In pratice, this means that we use information from old spikes to predict the sentiment at the current time window, what may be especially useful when the label is incorrectly predicted by the model we presented in Section 3.

4.1 Experimental Evaluation

We incorporate the textual feature vector $\mathcal{D}_t$ in a learning model by interpreting $P(\mathcal{Y}|\mathcal{S})$ estimates from Section 3 as *probabilistic labels* (or *soft labels*), which can then be incorporated into a variety of supervised learning algorithms [42, 36]. We have chosen to employ a version of Multinomial Naive Bayes extended to consider probabilistic labels [38]. We make this choice because of the easiness to extend Naive Bayes to incorporate probabilistic labels and its suitability for stream classification, since conditional term-class probabilities can be easily updated as more data is processed. Our features correspond to unigrams, bigrams and trigrams represented with term-arousal weights.

Figure 5 shows how accuracy varies, in the SOCCER dataset, as we vary the number of features we include in the model, considering both our term arousal representation and the traditional TF-IDF representation. We varied a threshold at the time window level, i.e., we included in the model the top K-ranked features on each time window. In addition to being more effective, the term arousal representation allows the sentiment model to be very compact, since the best accuracy were obtained by considering just the top 50 terms on each time window. Results are similar in the NFL dataset.

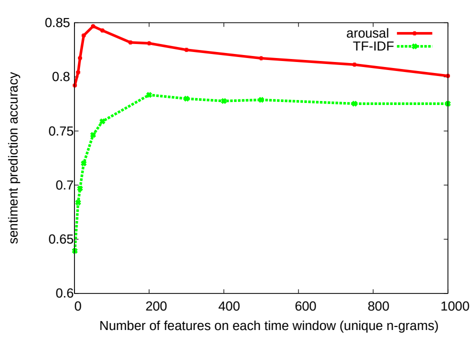

Figure 5: Accuracy vs top-K features comparing term-arousal and TF-IDF feature representation – SOCCER dataset.

In Figure 6 we show the increase on accuracy per volume bin, when adding textual features to the model. The increase on accuracy on lower-volume bins can be interpreted as the "transfer" of the reliable social context from spikes to the lower-volume time windows through the terms: when a high-arousal term is used on a low-volume time window, it contributes to the correct prediction of such time intervals.

4.2 Real-time sentiment tracking of live matches

To illustrate the usefulness and the utility of our combined label acquisiton/feature representation method, we now analyze the sen-

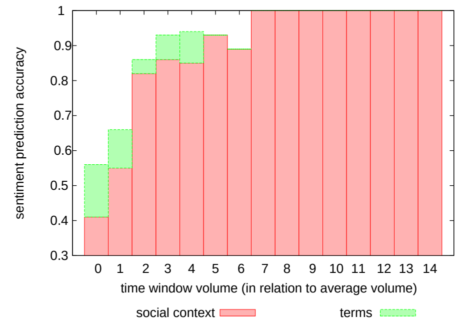

Figure 6: High-arousal n-grams carry the informative social contexts from the spikes to subsequent low-volume time windows – SOCCER dataset.

timent of the crowds expressed on Twitter during some interesting matches. For each match, we show the variation on the sentiment score over time in conjunction with the overall volume of tweets from each crowd. The scores are obtained by computing the ratios between the positive and negative probability estimates of the Naive Bayes classifier. Figure 7 shows the reactions of the supporters during SuperBowl 2011:

1. The Green Bay Packers score two touchdowns in the first quarter, reflected on the two spikes of happiness before 200'.
2. At 200' the Steelers scores a touchdown, and, after another touchdown at 240', the mood of Steelers' fans are better than Packers for a significant part of the match.
3. After a sequence of touchdowns from both teams between 320' and 350', the game comes to an end at 360' and Packers is proclaimed SuperBowl winners. Note that the majority of changes in the dominant sentiment of each crowd occur after a spike in the volume of messages, indicating that users are reacting to events. Note, also, that after the spike at 360' related to Packers' victory, our content-based classifier is capable of keeping track the positive sentiment towards Packers, in part because of high-arousal terms such as the ones shown in Table 2.

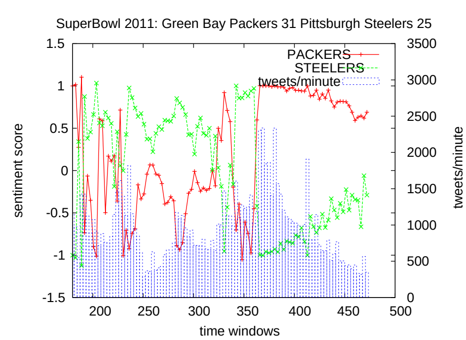

Figure 7: Sentiment variation during SuperBowl 2011 – Packers vs Steelers.

In the 2012 SuperBowl, played on February 5th, we also detected changes in crowd's humour, as shown in Figure 8:

1. The New York Giants started the game scoring 2-0 at 158' and 9-0 with a touchdown at 168'.

2. The Patriots scored two touchdowns in a row, at 224’ and 265’, reversing the expectations about the game outcome.

3. The Giants managed to score a touchdown in the last minute of the game and were proclaimed the 2012 SuperBowl champions at 298’, generating a long period of happiness on their supporters, whereas Patriots supporters were upset.

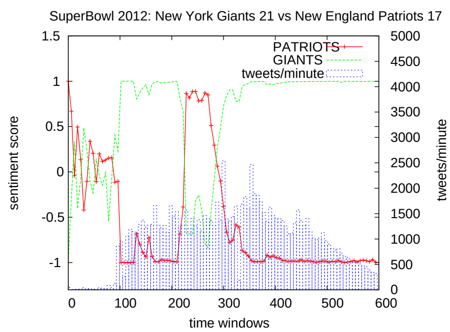

Figure 8: Sentiment variation during SuperBowl 2012 – Giants vs Patriots.

Soccer. We also illustrate our results with two matches of the last round of the 2011 Brazilian Soccer League. In Figure 9, team Cruzeiro comfortably beats his fierce rival Atletico by a surprising score of 6-1, scoring two goals in the early minutes of the match. Our model was able to correctly capture the positive reactions of Cruzeiro fans, and negative reactions of Atletico supporters. The second match, in Figure 10, showed a totally different pattern: Vasco and Flamengo played at the last round of the Brazilian 2011 Soccer League and Vasco needed to win in order to have any chance of winning the championship title:

1. At 149’, Vasco scored, and our algorithm detected a sudden burst of positive sentiments for Vasco and negative sentiments for Flamengo.

2. At minute 199’, however, Flamengo scores (note the spike in volume of tweets), vanishing any chances of Vasco winning the title. Our algorithm detected a sharp negative spike for Vasco in that moment. Even after conceding a goal, Vasco supporters were still upset, as expected; this illustrates the capacity of our algorithm in learning from spikes and using the learned term polarities on the subsequent time intervals.

3. Note that we have been able to track different supporters’ reactions, even during “similar” events: although Atletico scored against Cruzeiro at 220’, it was already losing by 5-0, what kept Cruzeiro supporters at a better mood. On the other hand, Flamengo’s tie goal against Vasco was a much more important one, and, even though Vasco was not losing the game, that goal vanished their chances of winning the title.

5. RELATED WORK

Social media data has been successfully used to detect real-world events such as disease outbreaks [9], earthquakes [41] and recurring events such as goals and touchdowns in sports matches [29]. Most of these researches are not focused on the deviation between self-reported data and real data; it is implicitly assumed that the number of users who decide to react and comment on the events

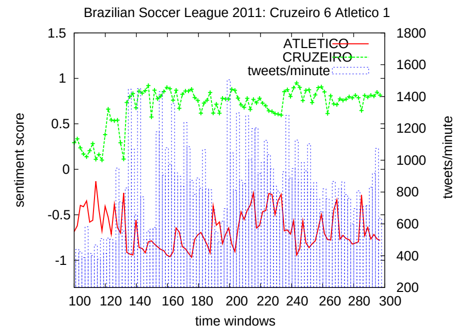

Figure 9: Sentiment variation during Brazilian Soccer League match – Cruzeiro vs. Atletico.

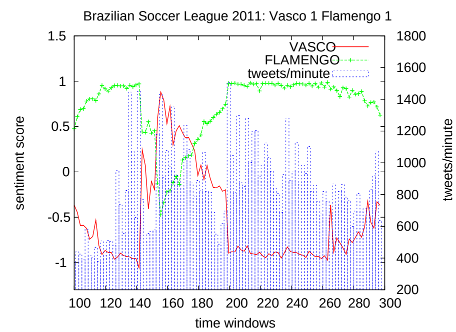

Figure 10: Sentiment variation during Brazilian Soccer League - Vasco vs. Flamengo.

being monitored will be large enough to allow detection. However, the self-reported nature of social media can strongly impact the observed social data, as observed by [28]: if we search in Twitter for the words “breathing” and “drinking water”, we may end up (wrongly) concluding that people usually drink more water than breath in their daily lifes. Some recent works try to compensate these biases in analysis of political debate, by observing that a small fraction of people intensively self-report their political opinions, while a silent majority does not [35], what can dramatically change conclusions and statistics on political behavior. Differently from these works, we stress that we aim to use self-reporting bias and the social/temporal contexts it creates to our benefit, in the design of better opinion analysis models, rather than correcting its effects.

Our work is closely related to research that explores opinion holder biases’ to perform sentiment analysis. Especially on the political domain, it is known that biases on opinion holders highly correlate to the type of opinion they express, and that social contexts based on groups of people with similar viewpoints provide useful signals for opinion analysis [20, 15, 31]. We add to these group-based social contexts a temporal perspective to explore the correlation between the real-world events taking place and the users currently reacting to what they are observing. To the best of our knowledge, this is the first attempt to detect positive and negative sentiment expressed on online media by capitalizing on the reasons that stimulate people to communicate more or less their feelings.

Sentiment analysis is still focused on static scenarios such as product reviews [37], on which lexicons of positive and negative words and traditional supervised machine learning techniques have been quite successful [45]. We are interested in sentiment analysis as a stream data mining task, a setting which requires learning algorithms to constantly update and refine data description models, in face of the time-changing characteristics of the data [14, 6]. The simultaneous presence of concept drift and lack of labeled data makes real-time sentiment analysis an even harder problem, since some standard solutions from one challenge make assumptions that do not hold in the other. The state-of-the art solution for coping with the scarcity of labeled data, *semi-supervised learning*, makes use of both labeled and unlabeled data for model generationand has also been applied to sentiment analysis [32]. However, due to the nonstationary characteristic of social streams, the usefulness of a few initially available labeled examples may be limited since they can become quickly outdated [13]. Conversely, the traditional approach for dealing with concept drift on nonstationary data is to incrementally update the model through fresh, recently-acquired labels that are provided by the stream [47], but this solution may not be feasible due to lack of labeled data. In terms of machine learning approaches, our algorithm is best related to *distant supervision* [18], which generate labeled data not by manual inspection of individual instances, but by applying some sort of heuristic/rule which output noisy labels. While distant supervision has considered emoticons as the source of labels, we take inspiration on social psychology patterns that guide people's reactions.

6. CONCLUSIONS

As a growing fraction of web content is generated in the form of social streams, there is a promising opportunity to build rich applications that track the emotional reactions of social media users during dynamically changing and potentially polarizing events such as sports matches, political debates and live breaking news. Traditional sentiment analysis, however, is not designed to operate on the stream setting, since the field has focused its attention on extracting opinions from static text such as product and movie reviews.

Real-time sentiment analysis is a difficult task; labeled data is usually not available to support supervised classifiers, and debate about monitored topics may turn into unpredictable discussions. We propose solutions to these challenges based on the different propensity users have on disclosing positive and extreme feelings, in comparison to negative and average feelings.

Since we mapped the usage of the social information on two machine-learning sub-tasks – acquisition of labeled data and feature representation – our work is orthogonal to current and future supervised models for real-time sentiment analysis. Depending on the characteristics of the domain and the social media platform, one or other sub-task may benefit more from our models.

We envisage a series of future research directions. In addition to experimenting our models on other domains (e.g, political debate and TV shows), we plan to enrich the social context we use to track sentiment by exploring the reaction patterns not only at group-level, but at user-level and on multi-group levels. At the user level, we can uncover different, more complex behavior of social media user posting patterns. Are there users which, in contrast to the dominant pattern, prefer to comment on negative experiences for their opposing sides than on positive events of their own side? At multi-group level, we may exploit the different relationships between polarized groups to generate more informative social contexts. For instance, supporters from rival teams are likely to follow and react whenever their rivals are being defeated, and that information could be embedded in the social context.

Another direction to improve our work is to better investigate the impact of time window sizes. In addition to automatically determine the optimal window size (or make it dynamic), analyzing effects of different window sizes in our models may unveil new patterns on how social media users react to real-world events.

Acknowledgments

This work was supported by InWeb - National Institute of Science and Technology for the Web, CNPq, Fapemig, and by Google, through its 2013 Google Brazil Focused Research Grants Program.

7. REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, New York, NY, USA, 2005. ACM.

[2] E. W. Anderson. Customer satisfaction and word of mouth. *Journal of Service Research*, 1(1):5–17, 1998.

[3] R. Balasubramanyan, W. W. Cohen, D. Pierce, and D. P. Redlawsk. Modeling polarizing topics: When do different political communities respond differently to the same news? In *ICWSM*. The AAAI Press, 2012.

[4] L. F. Barrett and J. A. Russell. The Structure of Current Affect: Controversies and Emerging Consensus. *Current Directions in Psychological Science*, 8(1):10–14, 1999.

[5] J. Berger. *Contagious: Why Things Catch On*. Simon & Schuster, 2013.

[6] A. Bifet and R. Kirkby. Data stream mining: a practical approach, August 2009.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[8] S. Brody and N. Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, EMNLP '11, Stroudsburg, PA, USA, 2011. ACL.

[9] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM.

[10] C. Dellarocas and R. Narayan. A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth. *Statistical Science*, 21(2):277–285, 2006.

[11] C. Dellarocas and C. A. Wood. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Manage. Sci.*, 54(3):460–476, Mar. 2008.

[12] E. Diener, R. Emmons, R. Larsen, and S. Griffin. The satisfaction with life scale. *J Pers Assess*, 49(1):71–5, 1985.

[13] K. B. Dyer and R. Polikar. Semi-supervised learning in initially labeled non-stationary environments with gradual drift. In *IJCNN*, pages 1–9. IEEE, 2012.

[14] J. a. Gama, R. Sebastião, and P. P. Rodrigues. Issues in evaluation of stream learning algorithms. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, New York, NY, USA, 2009. ACM.

[15] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. Konig. Blews: Using blogs to provide context for news articles. In *In Proceedings of the 2nd Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2008.

[16] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj. Limits of electoral predictions using twitter. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.

[17] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003.

[18] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford, 2009.

[19] P. H. C. Guerra, W. Meira, Jr, C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *7th Int'l AAAI Conf. on Weblogs and Social Media (ICWSM 2013)*, Boston, MA, 2013.

[20] P. H. C. Guerra, A. Veloso, W. Meira, Jr, and V. Almeida. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego, CA, 2011.

[21] Q. He, K. Chang, E.-P. Lim, and J. Zhang. Bursty feature representation for clustering text streams. In *SDM*. SIAM, 2007.

[22] N. Hu, J. Zhang, and P. A. Pavlou. Overcoming the j-shaped distribution of product reviews. *Commun. ACM*, 52(10):144–147, Oct. 2009.

[23] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, 2013.

[24] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60:2169–2188, 2009.

[25] A. H. Jordan, B. Monin, C. S. Dweck, B. J. Lovett, O. P. John, and J. J. Gross. Misery Has More Company Than People Think: Underestimating the Prevalence of Others' Negative Emotions. *Personality and Social Psychology Bulletin*, 37(1):120–135, Dec. 2010.

[26] I. Katakis, G. Tsoumakas, and I. Vlahavas. On the utility of incremental feature selection for the classification of textual data streams. In *10th Panhellenic Conference on Informatics (PCI 2005)*. Springer-Verlag, 2005.

[27] I. Katakis, G. Tsoumakas, and I. Vlahavas. Dynamic feature space and incremental feature selection for the classification of textual data streams. In *in ECML/PKDD-2006 International Workshop on Knowledge Discovery from Data Streams. 2006*. Springer Verlag, 2006.

[28] E. Kiciman. OMG, i have to tweet that! a study of factors that influence tweet rates. In *ICWSM*. The AAAI Press, 2012.

[29] J. Lanagan and A. F. Smeaton. Using twitter to detect and tag important events in live sports. *Artificial Intelligence*, 2011.

[30] R. Larson, M. Csikszentmihalyi, and R. Graef. Time alone in daily experience: Loneliness or renewal? *Loneliness: A sourcebook of current theory, research and therapy*, 1982.

[31] Y.-R. Lin, D. Margolin, B. Keegan, and D. Lazer. Voices of victory: a computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd int'l conference on World Wide Web*, WWW '13, 2013.

[32] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis digital library of engineering and computer science. Morgan & Claypool, 2012.

[33] M. M. Masud, C. Woolam, J. Gao, L. Khan, J. Han, K. W. Hamlen, and N. C. Oza. Facing the reality of data stream classification: coping with scarcity of labeled data. *Knowl. Inf. Syst.*, 33(1):213–244, 2011.

[34] D. Meshi, C. Morawetz, and H. R. Heekeren. Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use. *Frontiers in Human Neuroscience*, 7(439), 2013.

[35] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal minority versus silent majority: Discovering the opinions of the long tail. In *SocialCom/PASSAT*. IEEE, 2011.

[36] Q. Nguyen, H. Valizadegan, and M. Hauskrecht. Learning classification with auxiliary probabilistic information. In *Proc. of the 11th IEEE Int'l Conf. on Data Mining*, ICDM '11, Washington, DC, USA, 2011. IEEE Computer Society.

[37] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[38] G. Ramakrishnan, K. P. Chitrapura, R. Krishnapuram, and P. Bhattacharyya. A model for handling approximate, noisy or incomplete labeling in text classification. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, New York, NY, USA, 2005. ACM.

[39] M. Rost, L. Barkhuus, H. Cramer, and B. Brown. Representation and communication: challenges in interpreting large social media datasets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, New York, NY, USA, 2013. ACM.

[40] T. Ryan and S. Xenos. Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, 27(5):1658 – 1664, 2011.

[41] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, New York, NY, USA, 2010. ACM.

[42] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, New York, NY, USA, 2008. ACM.

[43] I. S. Silva, J. Gomide, A. Veloso, W. Meira, Jr., and R. Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *Proc. of the 34th Int'l Conf. on Research and development in Information Retrieval (SIGIR)*, New York, USA, 2011. ACM.

[44] P. Soucy and G. W. Mineau. Beyond tfidf weighting for text categorization in the vector space model. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, San Francisco, CA, USA, 2005.

[45] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*, 2002.

[46] G. Whannel. Reading the sports media audience. *MediaSport*, pages 221–232, 1998.

[47] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Mach. Learn.*, 23(1):69–101, Apr. 1996.

[48] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of the 14th Int'l Conference on Machine Learning (ICML)*, 1997.

[49] I. Zliobaite, A. Bifet, G. Holmes, and B. Pfahringer. MOA concept drift active learning strategies for streaming data. *Journal of Machine Learning Research*, 17:48–55, 2011.