ABSTRACT
A significant body of work focuses on detecting events in social media, much of which relies on restrictive assumptions like prior event-centric filtering or known language models. Rather than build these expensive pre-processing and filtering pipelines, we explore language-agnostic techniques using only temporal features like token frequency. We construct temporal features for tokens in an unfiltered Twitter stream, build a classifier to recognize tokens experiencing bursts in usage, and relate those bursts back to known occurrences within sports. Our results demonstrate preliminary feasibility in rapidly detecting tokens that correspond to high-impact occurrences within sporting events. We also successfully transfer a model trained on American football and hockey data to previously unseen types of sports like the 2014 World Cup and the 2014 Kentucky Derby/Belmont Stakes horse races. This model has less success at identifying events in regular-season Major League Baseball, so more research is required to enhance these capabilities.

Categories and Subject Descriptors
H.3.4 [Systems and Software]: Social Networks; H.3.3 [Information Search and Retrieval]: Information Filtering

General Terms
Experimentation, Performance, Verification

Keywords
transfer learning, event detection, sporting events, Twitter, social networks

1. INTRODUCTION
Social media’s new ubiquity has drastically altered our ability to spread and consume information, and researchers are now considering social networks as viable alternatives to real-time news sources. These social news systems are able to detect noteworthy events but often require restrictive assumptions on the target language, a priori knowledge of event types, pre-specified event-centric queries for filtering, retrospective analysis, and expensive language models for token processing. Resulting systems then become brittle and inflexible in application and are difficult to adapt to new domains or languages. While quickly detecting events without relying on these restrictions is a difficult task, doing so could support better event discovery and more rapid detection. Such enhanced, language-agnostic event detection has many practical applications, especially as applied to sports journalism and advertising. Sporting events occur often and in many different languages, stretching journalists’ ability to cover them; an automated approach could support these journalists and marketers in extracting game highlights or scheduling more effective advertisement placements, thereby increasing viewership and revenue.

We investigate this task by leveraging temporal features in unfiltered social media streams to identify event-related tokens that experience usage bursts, which might correspond to actual events. Our techniques learn an event detection classifier from these temporal features and a set of ground truth from selected events over the past five years. By focusing this classifier on events related to sporting competitions across a subset of sports types (e.g., American football and hockey) and applying the learned model to previously unseen sports types like association football, we also evaluate our model’s potential to adapt to different domains. The events on which we focus include a National Football League (NFL) championship game, Triple Crown horse races, the National Hockey League’s (NHL’s) Stanley Cup playoffs, a small set of Major League Baseball (MLB) games, and the FIFA World Cup to name a few. Given these data sources, the experiments outlined herein demonstrate the feasibility and possible limitations of identifying these bursting event-related tokens for both seen and unseen sporting event types.

2. RELATED WORK
With social media’s explosive popularity and the ease with which users can post information using a variety of means (mobile applications, the Web, or via text message), the huge volumes of data now being published has proven useful for a variety of purposes, the most popular of which concentrates on event detection and summarization. In 2009, researchers began transferring expertise on event detection from traditional news and blog data to social network-based microblogs like Twitter. Nagarajan et al. adapted exist-
ing spatio-temporal analysis to their Twitris framework to identify and localize specific themes according to region [6]. Twitris relies on Google’s “Insights for Search” to identify trending keywords for a given location or time interval; these keywords are fed into Twitter’s search API to bootstrap data collection. Twitris clusters these trending tokens to identify thematically similar content and present groups of related tokens as retrospective event “storylines.” Though Twitris and similar systems [10, 5] are powerful, their event summaries are too coarse to detect individual occurrences in sports and rely heavily on traditional media to bootstrap the detection process.

Cataldi et al. take a different approach in their 2010 paper on detecting emerging topics in Twitter by leveraging “user authority” as calculated using the well-known Google PageRank algorithm [2]. Though this approach might be useful for identifying authoritative sources like sports journalists or official team Twitter accounts, our approach goes user characteristics. Instead, we prioritize token-centric burst information over a user’s network influence since only a limited number of fans tweeting about some event may be authoritative in the network.

Work by Petrović, Osborne, et al. is perhaps the most similar system to ours [8, 9, 7]. This system, called ReDites, relies on locality sensitive hashing (LSH) to enable near real-time tweet clustering for event detection but is restricted to only for English-language tweets. LSH allows for fast similarity calculations to determine a message’s “nearest neighbor,” which enables high-speed clustering for theme/event generation. Once ReDites constructs these themes, it can perform a retrospective analysis to identify the first story related to a given theme. Our approach has a language flexibility that ReDites lacks, and we focus more on identifying events as they occur rather than retrospectively, but ReDites’s ability to handle many Twitter messages at an extremely high rate is an impressive benchmark for which we are striving for future versions.

Only recently has event detection specific to sports gathered more attention. Lanagan and Smeaton tried to align changes in tweet volume of a filtered Twitter stream with an annotated audio/video analysis and showed that bursts in Twitter usage co-occurred with high-impact events [3]. Zhao et al. used a lexicon of American football-related terms to refine the Twitter stream and detect events during the 2010-2011 NFL football season within 40 seconds [12]. Vasudevan et al. used Twitter to identify events specific to American football and found that events could be detected within a few minutes of the actual event [11]. The common threads among all these approaches, however, are prior knowledge of event type and a pre-specified set of event keywords, which limits their applicability to international sports and other languages.

3. METHODOLOGY

Our experiments took temporal features of tokens in an unfiltered Twitter stream and classified each token as bursty or non-bursty. We first discuss the set of temporal features we generated for each token, and then cover the algorithms we used for training our classifiers. From there, we briefly describe evaluation metrics for selecting high-performing models and techniques for boosting performance using semi-supervised learning and close with how we trained these models.

A high-quality feature for detecting bursts should yield higher weights for tokens that deviate significantly from normal behavior with respect to frequency or network density. For this research, we used the following features:

- **Frequency Regression** Given the log of a token’s frequency at each slice in the current window, take the slope of the best-fitting line.
- **Message Frequency Regression** Given the log of the number of tweets in which a token appears at each slice in the current window, take the slope of the best-fitting line.
- **User Frequency Regression** Given the log of the number of users using a token at each slice in the current window, take the slope of the best-fitting line.
- **Average Frequency Difference** Take the difference between the token’s frequency in the most recent slice and the average frequency across the current window.
- **Message Average Frequency Difference** Take the difference between the number of messages in which a token appears in the most recent slice and the average number of messages containing that token across the current window.
- **User Average Frequency Difference** Take the difference between the number of users who use a token in the most recent slice and the average number of users across the current window.
- **Density** Given a graph with nodes as users who use a given token and edges corresponding to mentions in tweets, calculate the graph’s density.
- **Entropy** Take the entropy of the set of tweets containing a given token.
- **TF-IDF** Calculate term frequency, inverse document frequency for a given token.
- **TF-PDF** A modified version of TF-IDF called term frequency, proportional document frequency [1].
- **BursT** Weight using a combination of a given token’s actual frequency and expected token frequency [4].

To support streaming data, we generated these features using a ten-minutes-wide sliding window. Each window was further comprised of three-minute slices such that each slice overlapped the following slice by two minutes. In addition, we normalized all features into a range of [0, 1] to avoid biases from scale in the support vector machines (SVMs) and random forests (RFs) we used for classification. These algorithms also have tunable parameters, and to find a good parameter set, we performed a grid search along the parameter space during learning and used the parameters that resulted in the highest F1 scores. To boost accuracy further, we used one step of the self-training heuristic to propagate positive labels to previously unlabeled tokens.

To support learning in these models, we first needed ground truth with known tokens and timestamps that correspond to actual events within sporting competitions. We reviewed collections of Twitter data already available or easily collectable and generated a series of timestamped events and
related tokens for thirteen different sporting events: the 2010 National Football championship game from a Twitter corpus from the University of Edinburgh, three soccer games in Nov. of 2012 from a collection of localized Argentinian tweets from Twitter’s firehose, and a collection of tweets from Twitter’s public data stream for a day’s worth of baseball games in June of 2014, the NHL’s 2014 playoffs, the NBA’s 2014 playoffs, the 2014 Kentucky Derby, the 2014 Belmont Stakes, and the 2014 World Cup. For each event, we used existing blog posts, news articles, and social media data to construct timelines for in-game events like scores, fumbles, and penalties, which became our positive data. We also required negative samples to differentiate between bursty and non-bursty tokens, but acquiring negative samples is difficult since we cannot know all tokens that should be bursting on Twitter at a specific time. We can, however, identify some tokens that should never be considered bursty: stop words; as such, we naively tagged English and Spanish stop words as negative samples.

4. EXPERIMENTAL SETUP

For the following experiments, we split our data into training and test sets such that events from each sporting match appeared in only one of the two sets, and the test set included some types of sporting matches that were not present in the training data (e.g., events from horse races and baseball appeared only in the test set). In this manner, we could evaluate our approach on two different problems: detecting events in previously seen types of sports, and detecting events in unseen types of sports. This first experiment used event data from NHL games, NBA games, and Argentinian soccer games to detect events during the first game of the 2014 World Cup, Brazil versus Croatia. In the second experiment, we applied this learned model to previously unseen sports types: the 2014 Kentucky Derby, the 2014 Belmont Stakes, and a small set of MLB games.

To determine accuracy, we applied our classifier to every window in a given sports match from five minutes before the match began to ten minutes after the match ended. Each window then generated a set of positively classified tokens that we searched for tokens related to known events like scores and penalties, and we plotted where these tokens were detected in relation to the actual event.

5. PRELIMINARY RESULTS

We trained the SVM and RF classifiers using a grid search for parameters on the training data as described and selected the highest-performing models with the best F1 scores. This training procedure yielded the results shown in Table 1.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Params</th>
<th>Prec.</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>c = 128, (\gamma = 16)</td>
<td>0.8723</td>
<td>0.5857</td>
<td>0.7009</td>
</tr>
<tr>
<td>RF</td>
<td>trees = 128, features = 9</td>
<td>0.9149</td>
<td>0.6418</td>
<td>0.7544</td>
</tr>
</tbody>
</table>

For the experiment on detecting events in the Brazil-Croatia World Cup game, we specified a set of tokens that should trend around issued penalty cards (“card”, “yellow”, and “red”) and goals (“goal”, “gol”, and “golazo”). In Figure 1, we show three in-game event types marked by colored bars, with game time in the horizontal axis and detectors on the vertical. The line marked “Actual” corresponds to the ground truth for this game, and the “SVM” and “RF” lines show where and for how long our classifiers identified relevant tokens as bursting. One can see our techniques detected all goal events and all but one of the penalty card events (we missed the last yellow card against Brazil’s Gustavo at 21:47 GMT) with a small number of false positives. For goals, our classifiers exhibited minimal detection delay, but that delay increased slightly for penalty cards.

For the two Triple Crown races, we identified the times in which the races started and specified the winning horses as target tokens; for the Kentucky Derby, these tokens were “california” and “chrome”, and for the Belmont Stakes, the target token was “tonalist”. As with the World Cup data, Figures 2 and 3 illustrate where our algorithm detected these known keywords in relation to the actual events. Once again, our classifiers seemed to detect event-related bursts appropriately.

An interesting side effect of this work was that our techniques were able to detect modifications of target tokens that also experienced bursts. For the World Cup games,
our techniques not only detected “goal” in English, Spanish, and Portuguese as expected but also modifications of those terms such as “goooaaaal” and “gooool” and similar iterations, which might otherwise have required expensive normalization techniques. Similarly for the Belmont Stakes race, we detected the winning horse’s name, “Tonalist,” but also saw a burst in the token “Totalist,” which is likely an autocorrected form of the horse’s name.

Despite these successes, the features and methods described were unable to identify in-game events in baseball games during a particular Friday evening in June of 2014. It seems no specific run- or team-related tokens experience bursts to the same degree as the other data we explored. In other words, none of the tokens our models classified as bursty seemed relevant to runs, outs, or other baseball-specific terminology, leading to no detected events.

6. CONCLUSIONS

To revisit our motivations, the goal for these experiments was to demonstrate the feasibility of detecting events and event-related tokens through analyzing temporal characteristics from unfiltered Twitter data streams. While many social media-based event detection systems require some form of query-based filtering and language model processing, our approach is more flexible, lighter weight, and easily adaptable to different domains. Based on experiments, we have demonstrated the utility of these temporal features in detecting events in both previously seen and unseen types of sporting competitions. At the same time, we encountered limitations regarding events with lower impact like scores or outs in regular-season MLB games, so additional research is necessary to explore methods for boosting such missed events. This preliminary ability to identify event-related tokens across seen and unseen event types, different languages, and without the need for normalization shows potential for detection methods that could be integrated into automated, near-real-time highlight generators or advertising auctions or leveraged in many other non-sports-related areas.

7. REFERENCES