[*reviews: Negatives:*
*Improvement in predictions not persuasive*
*prediction algorithms itself is very simple and naive*
*word clouds not informative..though takes up substantial space in paper*
*elections prediction is a naive purpose for DQE.. humans can easily define phrases for election*
*related works incomplete*
*comparison of election predictions to just static keywords is weak*
*compare with other static approaches-topic modelling*
*compare with state of art query expansion*
*use of social media for election prediction not novel..contribution not clear.*
*compare to other dqe approacehs and show our approach is better*
*Positives:*
*problem is real and important*
*continuous learning*
*number of elections used*
*approach is intuitive* ]

# Discovering Evolving Political Vocabulary in Social Media

Aravindan Mahendiran*, Bert Huang†, Wei Wang*,
Jaime Arredondo Sanchez Lira‡, Lise Getoor†, David Mares‡ and Naren Ramakrishnan*
*Virginia Tech.,
†University of Maryland
‡University of California, San Diego

[**TODO:** Add new experiments and latest results **Assignee:** Wei]
[**TODO:** Formatting once all new inputs are added
Restrict total size to 7 pages
Remove Authorlist for double-blind submission **Assignee:** Aravind]

***Abstract*—As a surrogate data source for many real-world phenomena, social media such as Twitter have yielded key insight into people's behavior and their group affiliations and memberships. As a phenomenon unfolds on Twitter, the language, hashtags, and vocabulary used to describe it evolves over time so that it is difficult to *a priori* capture the composition of a social group of interest using static keywords. Capturing such dynamic compositions is crucial to both understanding the true membership of social groups and in providing high-quality data for downstream applications such as trend forecasting. We propose a novel unsupervised learning algorithm that builds dynamic vocabularies using Probabilistic Soft Logic (PSL), a framework for probabilistic reasoning over relational domains. Using eight presidential elections from six countries of Latin America (Mexico, Venezuela, Ecuador, Paraguay, Chile, and Honduras), we show how our query expansion methodology helps capture dynamic trends and better ascertain group memberships of users. The ability to grow a vocabulary concomitantly with social media trends helps capture key milestones in election campaigns. In addition, we demonstrate that the results of such query expansion helps consistently improve the performance of election forecasting algorithms.**

## I. Introduction

It is now established that social media such as Twitter serves as a weak predictor or as a correlative surrogate for many real-world trends such as box office earnings [1], flu case counts [2], and even stock prices [3]. A topic of current interest is to study how online chatter can be used to model the social, economic or political landscape of a country.

Most approaches to tracking real-world phenomena over Twitter rely on first defining a vocabulary of keywords (or hashtags) to track over the social media. This is typically sufficient for phenomena about which we have a fairly stable understanding but not for rapidly evolving phenomena, e.g., an election season where new developments can upset or raise the success rates of competing candidates among key groups or populations. In such cases, rather than a static vocabulary to *a priori* define social groups of interest (e.g., Twitter users who are favorably or not favorably disposed toward specific candidates), it is preferable to grow dynamic vocabularies by modeling the temporal progression of events. In addition to better defining social group memberships, such modeling can provide higher-quality data for downstream applications such as forecasting.

We propose a dynamic query expansion strategy that aids in modeling social groups of interest over time, and use the domain of elections to demonstrate the effectiveness of our approach. Modeling elections provides both qualitative and quantitative insight into the utility of our approach. Our key contributions are:

1) We demonstrate a novel unsupervised learning algorithm that builds dynamic vocabularies using Probabilistic Soft Logic (PSL) [4], a framework for probabilistic reasoning over relational domains. Beginning with a small seed set of keywords/hashtags, we demonstrate how our PSL program helps grow the seed set into vocabularies involving hundreds of relevant terms. This aids in significantly improving the retrieval of tweets corresponding to relevant social groups of interest.
2) Using eight presidential elections from six countries of Latin America (Mexico, Venezuela, Ecuador, Paraguay, Chile, and Honduras), we show how our query expansion methodology helps capture dynamic trends and better ascertain group memberships of users. Our experiments over the Latin American region (where tweets arise from a multilingual set, including English, Spanish, Portugese, and French) demonstrates the practical utility of our approach.
3) In addition to qualitative understanding of social groups, we demonstrate how our approach yields appreciable improvements to the forecasting of election outcomes. By regressing against opinion polls in each region, we illustrate how better social group modeling reveals popularities of candidates and their policies.

## II. Related Work

Related work can be organized into key categories: query expansion, forecasting elections using social media, and probabilistic reasoning with PSL. [*reviews: refer
Dongsheng Duan, Yuhua Li, Ruixuan Li, Rui Zhang, and Aiming Wen. 2012. RankTopic: Ranking Based Topic Modeling. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM 12) [This work also captures relational information] ; Daniel Ramage, Susan T. Dumais,*

*and Daniel J. Liebling. ICWSM, The AAAI Press, (2010); Castella, Quim and Sutton, Charles A Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. CoRR abs/1301.0503 (2013)*
*Lau, Jey Han, Nigel Collier and Timothy Baldwin (2012) On-line Trend Analysis with Topic Models: hashtag twitter trends detection topic model online, In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)* ]

### A. Query Expansion

Query expansion is a classical technique in information retrieval (IR) [5] intended to improve retrieval performance by overcoming problems such as *synonymy*. Query expansion algorithms are typically iterative in nature wherein a seed set of query terms help identify an initial set of documents matching the query, and the highly-ranked retrieved documents (either judged by a human or by 'pseudo-relevance' techniques) are used to automatically grow the vocabulary. Modern query expansion algorithms use sophisticated concept modeling approaches [6] to grow the given seed set. In contrast to such classical approaches, our proposed approach is based on a dynamic query expansion strategy intended to track a vocabulary over time. Furthermore, unlike most approaches to query expansion, our PSL approach uses a probabilistic formalism to grow the vocabulary. Finally, PSL provides a rich programmatic environment to incorporate multiple indicators (social network, demographics, time) to grow the vocabulary rather than pre-committing to a specific strategy.

### B. Forecasting Elections using Social Media

Traditional approaches to forecasting elections use 'volume-based' ideas [7]–[10], i.e., forecasting election results by assessing the popularities of candidates and their policies. In particular, both [8], [10] fit a regression model to opinion polls with volume of mentions and sentiment as independent variables and the opinion polls as the dependent variable. More sophisticated approaches, as presented in [11]–[13], either model the candidates or the voters in the elections rather than compute the aggregated sentiment of the mass. In [12] the authors build a support vector machine (SVM) classifier trained on manually labeled tweets and classify users into 'left' and 'right' aligned. Using this information and how political information diffuses in a network, they demonstrate an accuracy of 95% in predicting the political alignment of Twitter users. Livne et al. in [11] analyze the Twitter profiles of candidates who contested the 2010 mid-term elections in the U.S. They identify topics specific to groups of candidates, split according to their known political orientations and use these features obtained as inputs to a regression model to forecast the elections. In a similar technique Diaz-Aviles in [13] model the candidates by building an emotional vector for each candidate using the mentions of that candidate and sentiments associated with each mention learned using the NRC Emotion Lexicon (EmoLex). They then use these profiles to predict the rise and fall of a candidate's popularity. In another thread of research, Mustafaraj et al. [14] model the distribution of political content among Twitter users. They divide the users into two groups, the "vocal minority" and the "silent majority" and observe that these two groups engage in different ways over social media. The vocal minority aims to broaden the impact of tweets by re-tweeting and linking to other web content, whereas the silent majority who tweet significantly less are more inclined to share their personal view points.

### C. Flawed Studies?

Recent coverage of election forecasting using Twitter has been critical, e.g., see [15], [16] These publications not only list pertinent problems in using Twitter to forecast elections but also detail recommendations on how to make such methodologies better. Gayo-Avello surveys the state-of-the-art approaches in predicting elections, most of which have been detailed above. Gayo-Avello argues that post-hoc analysis of elections in retrospect must not count as valid predictions and that researchers must be wary of the *file drawer* effect, i.e., the act of filing away negative results and publishing only the positive results. His major points of contention against such models are: lack of explainability, no direct way to model a 'vote' in social media, self-selection bias, unrepresentativeness of Twitter demographics, lack of sophisticated sentiment modeling strategies (e.g., to detect humor and sarcasm that abound in political conversations), and inability to capture indifferences among the voting public (i.e., abstaining from tweeting about politics can carry as much signal as explicit mentions of candidates). Finally, it has been shown that many of these models do not outperform a simple base model that forecasts success for the incumbent. Our approach here is to aid in better modeling of social groups and improve such predictions. A truly comprehensive system will utilize social media as just one of its strategies in forecasting and we do not make any claims of developing a universal forecasting system for elections.

### D. Social Group Modeling

Twitter is a natural venue to study social group modeling and there have been many studies that aim to study social groups in the context of election seasons. [17], for instance, models user affiliations within groups by capturing social networks comprising users, their posts, messages to other users and the various groups they intend to model. Dynamics of group affiliations are captured through various interactions between these entities in the underlying network. The authors discover hashtag usage specific to political seasons. Our work here differs in that we model the dynamic growth of vocabularies and, more importantly, use the resulting vocabularies for improving forecasting performance.

## III. Probabilistic Soft Logic

*Probabilistic soft logic* [4], [18] is a framework for collective probabilistic reasoning in relational domains. PSL models have been developed for various problem areas, including opinion diffusion [19], ontology alignment [18], trust in

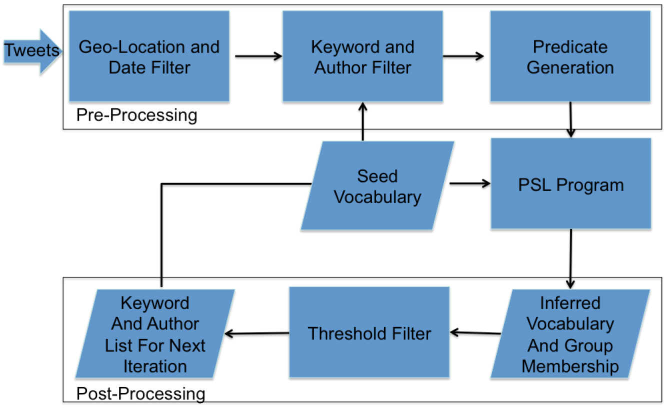

Tweets
Geo-Location and Date Filter
Keyword and Author Filter
Predicate Generation
Pre-Processing
Seed Vocabulary
PSL Program
Keyword And Author List For Next Iteration
Threshold Filter
Inferred Vocabulary And Group Membership
Post-Processing

Fig. 1: Design of the query expansion pipeline.

social networks [20], and knowledge graph identification [21]. PSL uses a syntax based on first-order logic to encode probabilistic models, which are declaratively defined as sets of weighted rules and constraints. PSL uses a continuous relaxation of logical truth to interpret these rules as a joint, continuous probability distribution over the truth values of logical atoms. The continuous relaxation enables fast algorithms to perform inference in highly structured models, as well as transparent incorporation of continuous variables and features. Like other rule-based systems, the ability to define complex models using a natural logical syntax streamlines the design process.

In PSL, user-defined *predicates* are used to encode the relationships and attributes, and *rules* capture the dependencies and constraints. Each rule's antecedent is a conjunction of atoms and its consequent is a disjunction. The rules are assigned non-negative weights, which correspond to the likelihood of the rules' satisfaction. The set of predicates and weighted rules thus make up a PSL program where known truth values of ground atoms are set from observed data and unknown truth values for the remaining atoms are inferred by finding a maximizing state of a probability distribution defined by the rules.

Given a set of atoms $\ell = \{\ell_1, \ldots, \ell_n\}$, an interpretation defined as $I : \ell \rightarrow [0,1]^n$ is a mapping from atoms to soft truth values. PSL defines a probability distribution over all such interpretations where those that satisfy more ground rules are more probable. The continuous interpretation of rule satisfaction uses the *Lukasiewicz t-norm* and its corresponding co-norm to define relaxations of the logical AND and OR respectively. Given interpretation $I$, these relaxations for the logical conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$) are as follows:

$$\begin{aligned} \ell_1 \,\tilde{\wedge}\, \ell_2 &= \max\{0, I(\ell_1) + I(\ell_2) - 1\}, \\ \ell_1 \,\tilde{\vee}\, \ell_2 &= \min\{I(\ell_1) + I(\ell_2), 1\}, \\ \tilde{\neg}\, l_1 &= 1 - I(\ell_1), \end{aligned}$$

where we use the tilde modifier (˜) to indicate the relaxation of the Boolean domain. Using the logical algebra and the relaxations above, an implication rule $r \equiv r_{body} \rightarrow r_{head}$ is satisfied if and only if the truth value of its head is at least that of its body. The rule's *distance to satisfaction* measures the degree to which this condition is violated.

$$d_r(I) = \max\{0, I(r_{body} - I(r_{head})\}$$

PSL induces a probability distribution over possible interpretations $I$ over the given set of ground atoms $l$ in the domain. Let $R$ be the set of all ground rules that are instances of a rule from the PSL program. The probability density function $f$ over $I$ is defined as

$$f(I) = \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \quad (1)$$

$$Z = \int_I \exp[-\sum_{r \in R} \lambda_r (d_r(I))^p] \quad (2)$$

where $\lambda_r$ is the weight of the rule $r$, $Z$ is a normalization constant, and $p \in \{1,2\}$ provides a choice between two different loss functions, linear and quadratic. The values of the atoms can be further restricted by providing linear equality and inequality constraints, allowing one to encode functional constraints from the domain.

Given a partial interpretation with grounded atoms based on observed evidence, *most probable explanation* (MPE) inference seeks the truth values for the unobserved atoms satisfying the most likely interpretation. The MPE inference is a convex optimization problem because the energy function is convex. This inference optimization can be efficiently solved using a fast decomposition algorithm [19], [22] by exploiting the structure of the energy function constructed from rules.

## IV. Dynamic Query Expansion using PSL

We use PSL to develop a dynamic query expansion strategy wherein we begin with an initial set of hashtags or terms (*seed words*) that we believe are indicative of the affinity of a particular user to a candidate contesting in the election (e.g., these terms are names, party symbols of the candidate). We iteratively use PSL inference over successive time windows such that the inference from window $w_t$ is used as a prior to window $w_{t+1}$, and the inference from that is used for window $w_{t+2}$, and so on. Figure 1 illustrates the design of the iterative algorithm for dynamic query expansion.

The initial pre-processing begins with the tweet input stream, which is filtered by a date range specified by the window size. For each election, tweets from a month leading up to the election were used. After extensive analysis we determined that the optimal window size was three days; smaller window sizes resulted in not enough data points for probabilistic inference, and larger window sizes lead to combinatorial explosion as PSL generates rules by substituting all possible groundings. Though the optimal window size could vary for different elections depending upon the number of tweets originating from the involved country, we use three days as window size for all elections for consistency. The tweets passing the date filter are then geocoded using a geolocation algorithm that infers the location of a tweet and enables us to localize the set of tweets analyzed. The geolocation algorithm tags the tweets with a location using the GPS coordinates of the tweet, if available, or landmarks and locations mentioned in the tweet or in the author's profile. For tweets that do not have any of these, we use a label propagation algorithm to infer the author's location through his/her network.

The geotagged tweets are then tracked for the presence of a hashtag from the vocabulary for that particular iteration. In addition to filtering tweets using the vocabulary the authors whose affiliations are already inferred by the system are also used as a filtering criteria. The information from the tweets are then coded into PSL predicates and fed into the inference process. The PSL program infers the hashtags and tweeters that are mostly associated with a particular candidate. Each author and hashtag's association with a candidate is measured using the truth value of the predicate grounding. In the post-processing step, these truth values are filtered by a threshold value to identify the hashtags and authors strongly associated to a candidate. These hashtags become a part of the vocabulary of the candidate and along with the users identified are used as a filter criterion for the next iteration. This iterative process proceeds until the day before the election when we obtain the final vocabulary which are strongly associated with a candidate.

Within the PSL program we define predicates to encode the network. The predicates $\textsc{Tweeted}(U,T)$ and $\textsc{Contains}(T,W)$ capture the fact that a user $U$ tweeted a tweet $T$ and tweet $T$ contains hashtag $W$ respectively. Similarly, the belief that an user $U$ or hashtag $W$ is affiliated/associated to the group $G$ is encoded as $\textsc{IsMember}(U,G)$ and $\textsc{Belongs}(W,G)$ respectively. In order to capture the temporal connectivity between the iterations, in addition to the initiating the inference process with the rule

$$\textsc{SeedWord}(W,G) \Rightarrow \textsc{Belongs}(W,G)$$

we define additional rules such as

$$\textsc{WasMember}(A,G) \Rightarrow \textsc{IsMember}(A,G)$$

$$\textsc{Belonged}(W,G) \Rightarrow \textsc{Belongs}(W,G)$$

where the predicates $\textsc{WasMember}$ and $\textsc{Belonged}$ are inferences from the previous time window and are loaded in as priors for the current iteration. These rules are weighted slightly lower than the recursive rules below so that the system overcomes the bias it had learned in light of new, more convincing evidence. This way hashtags that are more indicative of a user's affiliation are assigned stronger truth values or weights for every successive iteration and the truth values of hashtags that are not are reduced. The same reasoning applies to the user-candidate affiliations (memberships). Below we outline the recursive PSL rules that grows the hashtag preferences and the user affiliations.

$$\begin{aligned}\textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{Contains}(T,W) \tilde{\wedge} \textsc{Belongs}(W,G)\\ \tilde{\wedge} \textsc{Positive}(T) \Rightarrow \textsc{IsMember}(A,G)\end{aligned}$$

$$\begin{aligned}\textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{Contains}(T,W) \tilde{\wedge} \textsc{Belongs}(W,G)\\ \tilde{\wedge} \textsc{Negative}(T) \Rightarrow\sim \textsc{IsMember}(A,G)\end{aligned}$$

$$\begin{aligned}\textsc{IsMember}(A,G) \tilde{\wedge} \textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{Contains}(T,W)\\ \tilde{\wedge} \textsc{Positive}(T) \Rightarrow \textsc{Belongs}(W,G)\end{aligned}$$

$$\begin{aligned}\textsc{IsMember}(A,G) \tilde{\wedge} \textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{Contains}(T,W)\\ \tilde{\wedge} \textsc{Negative}(T) \Rightarrow\sim \textsc{Belongs}(W,G)\end{aligned}$$

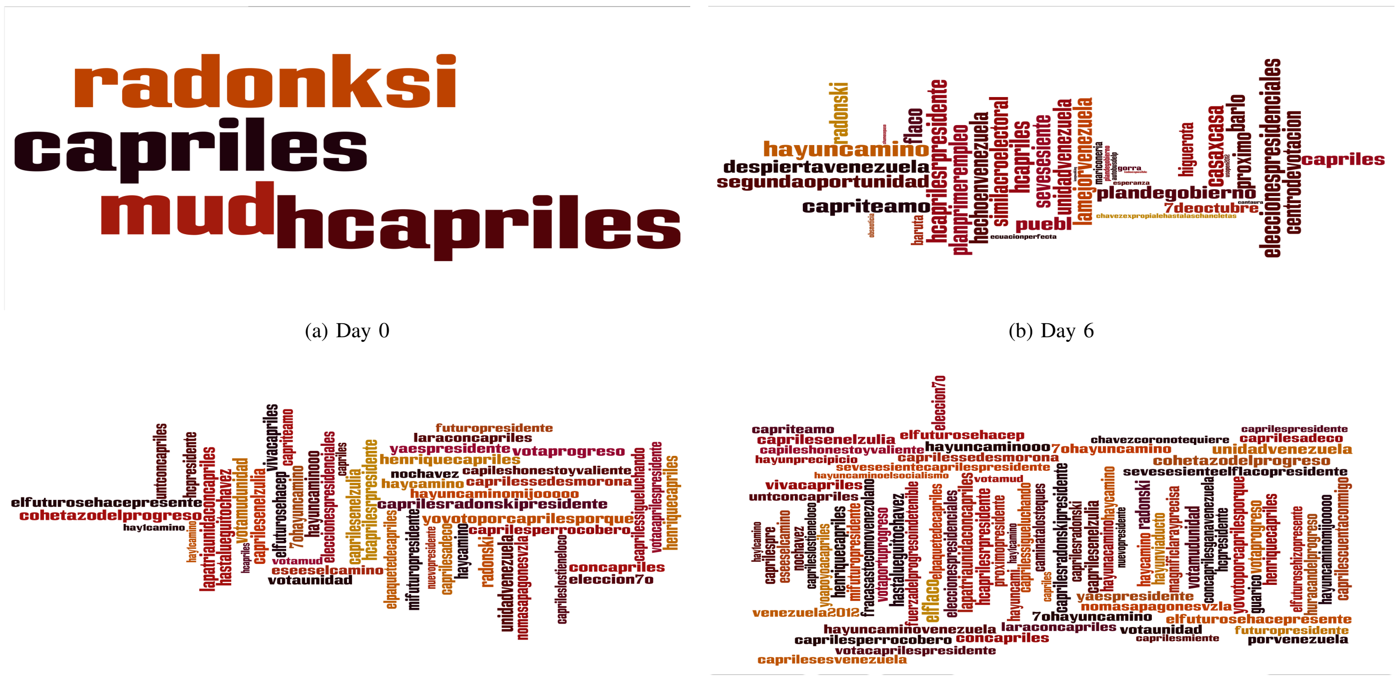
Here $\textsc{Positive}$ and $\textsc{Negative}$ are predicates whose truth values are calculated from the sentiment of the tweet such that the highly positive tweets get a truth value closer to 1.0 for the predicate $\textsc{Positive}$ and highly negative tweets are assigned a truth value of 1.0 for the predicate $\textsc{Negative}$. Since PSL works under the closed world assumption, we do not need to specify the groundings that are false i.e., positive tweets are not assigned 0.0 for the predicate $\textsc{Negative}$ and vice-versa. For tweets that do not have a positive or negative orientation we assign a truth value of 0.5 for both the $\textsc{Positive}$ and $\textsc{Negative}$ predicates.

We also defined rules that encode how ideologies propagate in a social media, specifically Twitter. For example, the first rule below states if one author retweets a tweet created by another author then it can be assumed that the former is endorsing the opinion of the latter and hence likely to have the same political affiliation. Similarly a person mentioning another person in a positive connotation is assumed to share similar views. The rules below detail the propagation of affiliation based on social interactions.

$$\begin{aligned}\textsc{IsMember}(B,G) \tilde{\wedge} \textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{ReTweet}(T,B)\\ \Rightarrow \textsc{IsMember}(A,G)\end{aligned}$$

$$\begin{aligned}\textsc{IsMember}(B,G) \tilde{\wedge} \textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{Mentions}(T,B)\\ \tilde{\wedge} \textsc{Positive}(T) \Rightarrow \textsc{IsMember}(A,G)\end{aligned}$$

$$\begin{aligned}\textsc{IsMember}(B,G) \tilde{\wedge} \textsc{Tweeted}(A,T) \tilde{\wedge} \textsc{Mentions}(T,B)\\ \tilde{\wedge} \textsc{Negative}(T) \Rightarrow\sim \textsc{IsMember}(A,G)\end{aligned}$$



(a) Day 0

(b) Day 6

(c) Day 15

(d) Day 30

Fig. 2: Evolution of hashtags for Henrique Capriles

The last two rules defined below encode the assumption that when two hashtags co-occur and one is a name of a candidate then the other hashtag is bound to be about the candidate too. Since these two rules have two variables W1 and W2, the number for rules generated by substituting actual groundings (hashtags) increase rapidly with the number of tweets feeding into the inference process. This was a major contributing factor to the memory issues detailed in the optimal window size discussion.

$$\begin{aligned}\textsc{Contains}(T, W1) \tilde{\wedge} \textsc{Contains}(T, W2) \tilde{\wedge}\\ \textsc{SeedWord}(W1, G) \tilde{\wedge} \textsc{Positive}(T)\\ \Rightarrow \textsc{Belongs}(W2, G)\end{aligned}$$

$$\begin{aligned}\textsc{Contains}(T, W1) \tilde{\wedge} \textsc{Contains}(T, W2) \tilde{\wedge}\\ \textsc{SeedWord}(W1, G) \tilde{\wedge} \textsc{Negative}(T)\\ \Rightarrow\sim \textsc{Belongs}(W2, G)\end{aligned}$$

In addition to the rules, we also define constraints on the BELONGS and ISMEMBER predicates so that a particular hashtag or author can be associated to at most one candidate. Once all the tweets are loaded into the PSL program as predicates, we start the inference process by closing all the predicates except ISMEMBER and BELONGS. This way, only the truth values of these two predicates are inferred and the other groundings of the closed predicates are regarded as facts.

## V. EXPERIMENTAL RESULTS

Our experimental results are organized alongside the following questions:

- How adept is our PSL-based dynamic query expansion algorithm at extracting relevant hashtags/keywords over the course of an election season?
- How does the performance of forecasting algorithms improve using our expanded vocabularies?

We answer each of these questions next.

### A. Election vocabularies inferred

**Venezuela:** Figure 2 shows how the hashtags for Henrique Capriles evolved during the month leading up to the election. Initially in Figure 2a the system begins with only a few hand picked hashtags that constitute the seed vocabulary. After a few iterations Figure 2b shows how the vocabulary has grown. However, not all the words identified until now remain in the final vocabulary as the system drops certain words in successive iterations. At the same time it is also noticed that hashtags like “capriles” and “hayuncamino” which are very strongly associated with Capriles consistently remain as the top ranked hashtags even after ten iterations (Figure 2d). It is also interesting to note that the algorithm identified hashtags like “nochavez” (Figure2c) and attributed it rightly to Hugo Chávez’s primary contender, i.e., Capriles. In Figure 3, the first plot elucidates how hashtags like *“elmnduconchavez”* and *“univistaconchavez”* remain highly associated with Hugo Chávez for the October 7th Presidential election. These hashtags remain indicative of a user’s affiliation throughout the month leading up to the election. Meanwhile hashtags such as *”beatles”* and *”facebook”* (in second plot) show spikes in their time series primarily because users affiliated with Chávez used them during that time window. But as the iterative

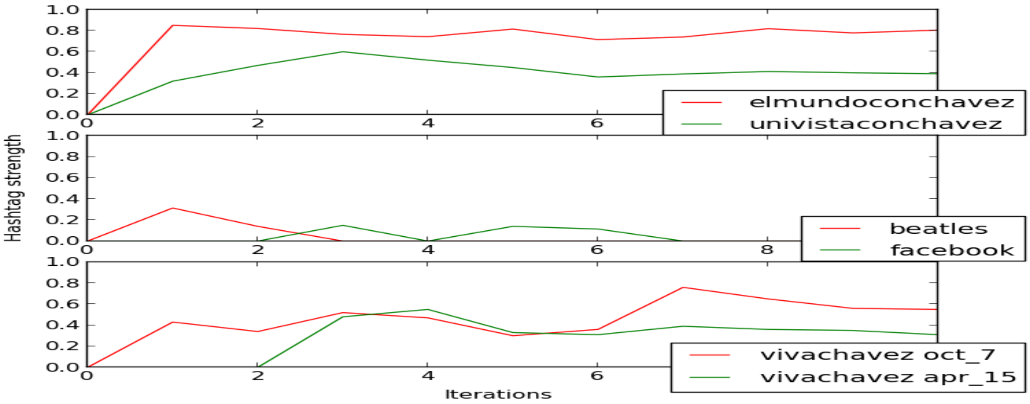


Fig. 3: Time series comparison for different hashtags identified for Hugo Chávez.

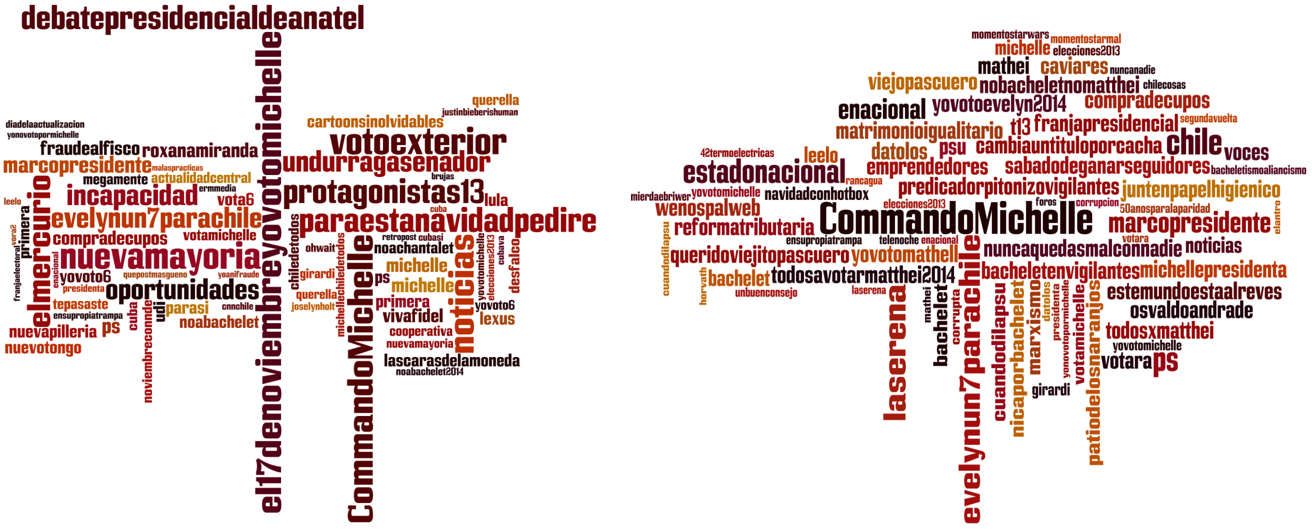
process continues, the system drops these non-informative words. The third plot presents another interesting observation. Hugo Chávez who had won the election on October 7, 2012 was diagnosed with cancer and passed away before being sworn in as the President. This triggered a re-election on April 15, 2013 where Nicolas Maduro, who had assumed the role of acting president then, competed against Henrique Capriles in the Presidential race. The hashtag *"vivachavez"* is part of both the elections, despite the fact that Hugo Chávez did not compete in the second election. It is picked up as a phrase commonly used by supporters of Nicholas Maduro whose election campaign was strategized around the death of Hugo Chávez to garner sympathy and mobilize support. Similarly variations of the hashtags *"hayuncamino"* and *"unidadvenuzela"* were returned for Henrique Capriles for both these elections. The tag MUD is for "Mesa de la Unidad Democratica" (Democratic Unity Roundtable) that was the organization created for the opposition to Chávez. The vocabulary grows to include other terms associated to the campaign, as the official slogan for the opposition "hayuncamino" (there is a road). Others relate to programs that Capriles wanted to implement, such as "planprimerempleo" (First Job Program).

**Mexico:** A general election in Mexico took place on July 1st, 2012. The two front runners where Enrique Peña Nieto (EPN) and Andres Manuel López Obrador (AMLO). The tags (Figure 5a) show the contest between these two candidates, the first belonging to the "Partido Institucional Revolucionario" (PRI). Among the tags we can also find reference to the "yosoy132" student movement that became a key player during the election. We also see "niunvotoalpeje" (no one vote for AMLO) and "soyantipri" (I am against PRI) which are attributed against AMLO and EPN respectively. **Paraguay:** Figure 5b details the election between Horacio Cartes from the Partido Colorado and Efrain Alegre from Alianza Paraguay. The incumbent president belonged to Partido Colorado and we can see some tags talking about a protest vote: "votocastigoya" (protest vote now). Also references from their campaign to each candidate as "yovotoporefrain" (I vote for efrain) or "todosconcartesavanzapais" (everyone with cartes, the country goes forward) are seen.

**Honduras:** In November 2013 Honduras had a general election to choose President, Congress and local officials. The wife of former President Zelaya, Xiomara Castro, contended against Juan Orlando Hernandez from the incumbent Partido Nacional. Both candidates showed similar numbers in the polls before the election. We can find tags that either support Xiomara, like "xiomarapresidenta" (Xiomara President), "hondurastienepresidenta" (Honduras has a female president) or against her, "noaxiomara" (no to Xiomara) in Figure 5c.

**Ecuador:** Ecuador had a first round for the presidential election on February 17th, 2013 where Rafael Correa obtained more than 50% of the votes needed to avoid a second round. From the tags(Figure 5d) we can read support in phrases such as "yatenemospresidente" (we have elected president) or "tenemosarafael" (we have Rafael). There are some references to other candidates such as "lassopresidente" (candidate Lasso president).

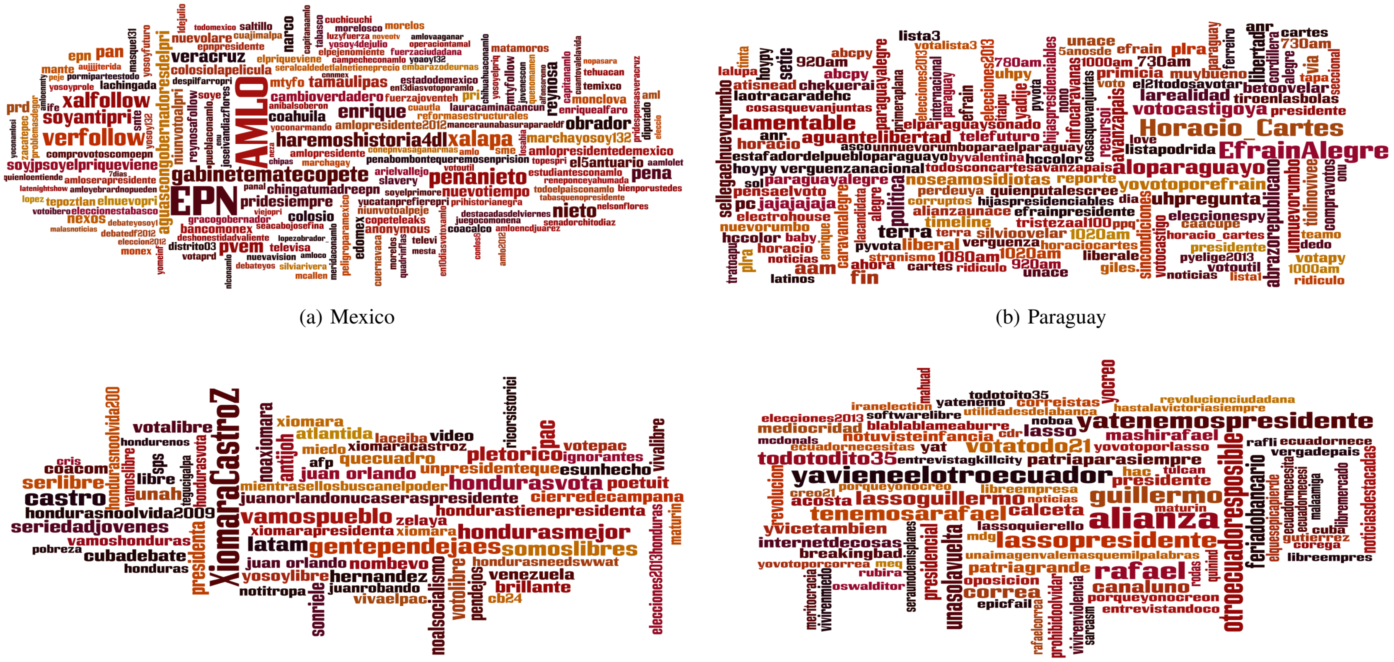
**Chile:** Figure 4 shows the hashtags identified for Michelle Bachelet who won the Chilean Presidential elections that was decided over two rounds. The first round was conducted on the 24th of November, 2013 and the 2nd round was conducted on December 15, 2013. The query expansion pipeline for Bachelet's group in both these elections were initialized only with three seed words: *Bachelet, CommandoMichelle and PS*. The first name of the candidate was not used as it introduced a lot of noise because "Michelle" is a very common name. The first figure shows the hashtags identified for Bachelet during the first round and the second figure for the second round. It can be seen that there is a lot of overlap in the vocabulary which is the expected outcome. Similarly there were a lot of common hashtags between the two rounds of election for the other candidate, Evelyn Matthei, too. It can be noticed that the vocabulary for the Honduran and Ecuadorean elections are quite noisy. This is primarily because Twitter is not as popular in these two countries as in Venezuela or Chile and therefore the number of tweets used for the inference was significantly



(a) Nov 24

(b) Dec 15

Fig. 4: Hashtags identified for Michelle Bachelet



(a) Mexico

(b) Paraguay

(c) Honduras

(d) Ecuador

Fig. 5: Vocabulary of hashtags identified for different elections

| Feature | Coefficient Value |
|---|---|
| SoPU | 0.4622 |
| SoNU | -0.443 |
| SoPM | 0.1158 |
| SoNM | -0.065 |
| SoS | 0.156 |
| Incumbency | 0.0 |

TABLE I: Regression coefficients learned for features; SoPU:Share of positive users;SoNU:Share of negative users;SoPM:Share of positive mentions;SoNM:Share of negative mentions;Sos:Share of sentiment;Incumbency:Binary variable for incumbency

lesser. This in turn affects the quality of the PSL inference.

### B. Forecasting Algorithms

We adapt two basic forecasting algorithms from current literature to evaluate the performance of our expanded vocabularies. The first model is a naive model, dubbed the *"unique visitor model"*, adapted from [9] and [7], and forecasts elections based on the counts of mentions of a candidate. The second model uses a regression fit to map from tweet features to opinion polls and thus forecast elections. We dub this model as the *"regression model"* and this approach is adapted from [10] and [8].

**Unique Visitor Model:** The assumption here is that large parties that are more popular will have a larger social media footprint than smaller and less popular parties. Further bot-generated tweets from election campaigns can artifically boost candidate mentions. This model estimates the relative popularity of candidates contesting the election by accounting for the unique users tweeting about the candidate and thus aiming to normalize such effects. The seed vocabulary for PSL is crafted by hand and includes the candidate's names and aliases, the name and acronyms for his/her political party and the official Twitter handle of the candidate. For the given time period, the tweets from the country in question are tracked for the occurrence of the terms in the vocabulary. We then build a time series of sentiment (from Lexalytics) and Klout scores (from Klout.com) from the tweets returned. The absolute popularity of a candidate $C_d$ is calculated as:

$$C_d = \sum_{i \in \text{Users}} K_i \times UCS_{id} \tag{3}$$

where $K_i$ is the Klout score for user $i$, and $UCS_{id}$ is User Candidate Score, the average of sentiment scores for all tweets from user $i$ about candidate $d$. These popularities are normalized and used to forecast the relative proportion of votes.

**Regression Model:** In this model, in addition to tweets, we leverage any opinion polls available for the elections to make our predictions. Like the earlier model we track the tweets that mention words/phrases from the vocabulary defined for each candidate. We then conduct a linear regression fit that uses the opinion polls as dependent variable and features generated from these tweets as independent variable. We reason that by regressing from the Twitter features to the opinion polls the bias due to Twitter being a non-representative sample can be mitigated. We use a total of six features: Klout scores, number of unique users, total number of mentions, sentiment, and incumbency. The Klout scores, unique users, and mentions are further categorized into positive and negative mentions. We normalize each of these features across all candidates to obtain the relative share of the volume. When there is more than one polling house publishing an opinion poll (for the same date) we take the average of the polls. Table I details the coefficients learned for each feature averaged over all the candidates from all the elections. The values confirm our hypothesis that the number of unique users and sentiment have more predictive power than total number of mentions. Intuitively it is also seen that the coefficients for share of negative users and negative mentions carry a negative weight. Another interesting observation is the fact that the incumbency binary variable is not as predictive which could be due to a confounding effect with other features. Having learnt a regression fit, we forecast the election by using features in the 10 day window leading upto the prediction date.

**Evaluation:** To evaluate our hypothesis about vocabularies we test our models on eight different presidential elections from Latin America using both the seed vocabulary and the vocabulary generated by the query expansion algorithm. To maintain consistency, we track only the hashtags identified by the query expansion pipeline until that particular date (i.e., newly inferred hashtags are not used to retroactively analyze older data). Figure 6 shows the increase in the number of documents that were used by the algorithms to make a predictions. Note that when averaged across all the eight elections we obtain close to a two-fold increase in the number of tweets that were used by these models. This is a substantial increase of relevant tweets for the domain.

To further illustrate the fact that the vocabulary used by such algorithms plays a vital role, we compare the performance of the models using the two different vocabularies. To reduce the effect of outliers we track the popularity of only the top two candidates from each election. Table II shows the reduction in prediction error for the Unique Visitor model for each candidate if the expanded vocabulary from the PSL approach is used instead of the seed vocabulary. On average the error was reduced by a 28.60% from the original prediction error obtained by using seed vocabulary. Similarly Table III shows the reduction in error for the Regression Model. Here an even better improvement of 45.19% reduction in error was noted. Averaging the reduction in error for both the models, the query expansion exercise was able to reduce the prediction error by 36.90%. We see greater and more consistent improvement with the regression model as the model weighs each window of tweets differently depending upon the opinion poll time series whereas the unique visitor model values them equally. Therefore, when the algorithm uses the 'not-so-informative' hashtags identified during the earlier iterations, the sentiment value and the counts of these mentions bring down the accuracy of the model even though at a later stage hashtags that are more strongly indicative of a user's preference is picked up. For instance, words such as "facebook" which occur commonly dominate the counts and therefore skew the results even though they are dropped from the vocabulary at a later point.

## VI. Discussion

We have demonstrated a novel query expansion methodology using PSL and illustrated how a suitably inferred vocabulary can provide significant gains in forecasting performance for elections. Although we have used elections to demonstrate performance gains, we believe that the dynamic query expansion system is general and can be applied to more domains.

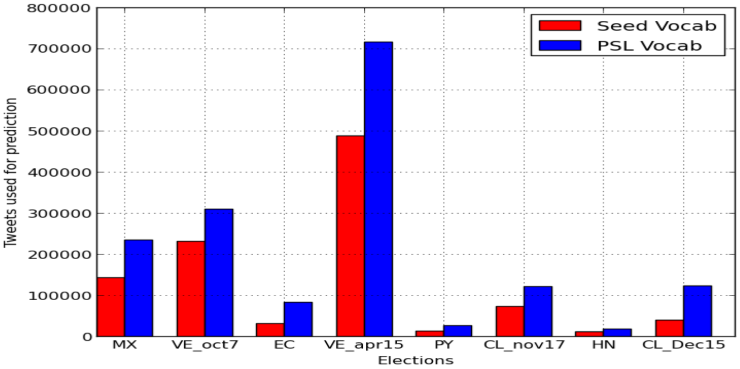


Fig. 6: Retrieval increases using the PSL vocabulary w.r.t. the seed vocabulary.

| Election | Candidate | Actual Result | Seed Vocab. | Error | PSL Vocab. | Error |
|---|---|---|---|---|---|---|
| Mexico | Peña Nieto | 38.1 | 49.26 | 11.11 | 46.43 | **8.28** |
| | López Obrador | 31.64 | 25.11 | 6.53 | 27.65 | **4.00** |
| Venezuela_Oct7 | Hugo Chávez | 55.07 | 63.69 | 8.62 | 55.24 | **0.17** |
| | Henrique Capriles | 44.31 | 36.31 | 8.00 | 44.76 | **0.45** |
| Ecuador | Rafael Correa | 57.16 | 32.36 | 24.81 | 32.90 | **24.27** |
| | Guillermo Lasso | 22.68 | 36.93 | **14.25** | 37.88 | 15.20 |
| Venezuela_Apr15 | Nicolás Maduro | 50.61 | 42.08 | 8.53 | 44.05 | **6.56** |
| | Henrique Capriles | 49.12 | 37.98 | **11.14** | 37.14 | 11.98 |
| Paraguay | Horacio Cartes | 48.48 | 29.80 | **18.68** | 29.12 | 19.36 |
| | Efraín Alegre | 39.05 | 27.21 | **11.84** | 26.63 | 12.42 |
| Chile_Nov17 | Michelle Bachelet | 46.70 | 26.62 | 20.08 | 29.92 | **16.78** |
| | Evelyn Matthei | 25.03 | 18.76 | 6.27 | 19.52 | **5.51** |
| Honduras | Orlando Hernández | 36.80 | 28.94 | 7.86 | 34.74 | **2.06** |
| | Xiomara Castro | 28.70 | 9.67 | 19.03 | 14.20 | **14.50** |
| Chile_Dec15 | Michelle Bachelet | 62.16 | 57.66 | 4.50 | 59.24 | **2.92** |
| | Evelyn Matthei | 37.83 | 42.34 | 4.51 | 40.67 | **2.84** |

TABLE II: Reduction in prediction error for Unique Visitor Model. All values shown are percentages.

| Election | Candidate | Actual Result | Seed Vocab. | Error | PSL Vocab. | Error |
|---|---|---|---|---|---|---|
| Mexico | Peña Nieto | 38.1 | 46.80 | 8.65 | 39.00 | **0.85** |
| | López Obrador | 31.64 | 24.67 | 6.97 | 28.64 | **3.00** |
| Venezuela_Oct7 | Hugo Chávez | 55.07 | 49.89 | 5.18 | 55.89 | **0.82** |
| | Henrique Capriles | 44.31 | 36.31 | 8.00 | 43.91 | **0.40** |
| Ecuador | Rafael Correa | 57.16 | 53.33 | 3.84 | 54.33 | **2.84** |
| | Guillermo Lasso | 22.68 | 12.27 | 10.41 | 12.75 | **9.93** |
| Venezuela_Apr15 | Nicolás Maduro | 50.61 | 51.45 | 0.84 | 50.58 | **0.03** |
| | Henrique Capriles | 49.12 | 35.96 | 13.16 | 38.11 | **11.01** |
| Paraguay | Horacio Cartes | 48.48 | 35.21 | 13.27 | 40.63 | **7.85** |
| | Efraín Alegre | 39.05 | 31.33 | 7.72 | 34.44 | **4.62** |
| Chile_Nov17 | Michelle Bachelet | 46.70 | 38.91 | 7.79 | 41.80 | **4.91** |
| | Evelyn Matthei | 25.03 | 19.20 | 5.83 | 20.98 | **4.05** |
| Honduras | Orlando Hernández | 36.80 | 25.16 | 11.64 | 28.30 | **8.50** |
| | Xiomara Castro | 28.70 | 16.53 | 12.17 | 24.90 | **3.80** |
| Chile_Dec15 | Michelle Bachelet | 62.16 | 39.12 | 23.04 | 39.80 | **22.37** |
| | Evelyn Matthei | 37.83 | 20.88 | 16.95 | 21.68 | **16.15** |

TABLE III: Reduction in prediction error for Regression Model. All values shown are percentages.

In future work, we aim to more finely model information about electoral demographics and study interactions both at the group and at the individual level. We also intend to use labeled data to learn PSL programs (both structure and probabilities). Finally, we aim to use the framework presented here as a platform to investigate theories of how social groups participate and influence elections.

## References

[1] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 492–499.

[2] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 115–122.

[3] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[4] A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor, "A short introduction to probabilistic soft logic," in *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012, pp. 1–4.

[5] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 1.

[6] D. Metzler and W. B. Croft, "Latent concept expansion using markov random fields," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 311–318.

[7] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, pp. 178–185, 2010.

[8] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM*, vol. 11, pp. 122–129, 2010.

[9] D. Saez-Trumper, W. Meira, and V. Almeida, "From total hits to unique visitors model for elections forecasting," in *International Conference on Web Science*, 2011.

[10] A. Bermingham and A. F. Smeaton, "On using twitter to monitor political sentiment and predict election results," 2011.

[11] A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic, "The party is over here: Structure and content in the 2010 election." in *ICWSM*, 2011.

[12] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*. IEEE, 2011, pp. 192–199.

[13] E. Diaz-Aviles, C. Orellana-Rodriguez, and W. Nejdl, "Taking the pulse of political emotions in latin america based on social web streams," in *Web Congress (LA-WEB), 2012 Eighth Latin American*. IEEE, 2012, pp. 40–47.

[14] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opionions of the long tail," in *Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (socialcom)*. IEEE, 2011, pp. 103–110.

[15] P. T. Metaxas, E. Mustafaraj, and D. Gayo-Avello, "How (not) to predict elections," in *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*. IEEE, 2011, pp. 165–171.

[16] D. Gayo-Avello, "" i wanted to predict elections with twitter and all i got was this lousy paper"–a balanced survey on election prediction using twitter data," *arXiv preprint arXiv:1204.6441*, 2012.

[17] B. Huang, S. H. Bach, E. Norris, J. Pujara, and L. Getoor, "Social group modeling with probabilistic soft logic," in *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models, and Applications*, 2012.

[18] M. Broecheler, L. Mihalkova, and L. Getoor, "Probabilistic similarity logic," in *Uncertainty in Artificial Intelligence (UAI)*, 2010.

[19] S. Bach, M. Broecheler, L. Getoor, and D. O'leary, "Scaling mpe inference for constrained continuous markov random fields with consensus optimization," in *Advances in Neural Information Processing Systems*, 2012, pp. 2663–2671.

[20] B. Huang, A. Kimmig, L. Getoor, and J. Golbeck, "A flexible framework for probabilistic models of social trust," in *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP)*, 2013.

[21] J. Pujara, H. Miao, L. Getoor, and W. Cohen, "Knowledge graph identification," in *International Semantic Web Conference (ISWC)*, 2013.

[22] S. H. Bach, B. Huang, B. London, and L. Getoor, "Hinge-loss Markov random fields: Convex inference for structured prediction," in *Uncertainty in Artificial Intelligence (UAI)*, 2013.