Discovering Evolving Political Vocabulary in Social Media

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

***Abstract*—As a surrogate data source for many real-world phenomena, social media such as Twitter can yield key insight into people's behavior and their group affiliations and memberships. As an event 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 10 presidential elections from eight countries of Latin America (Mexico, Venezuela, Ecuador, Paraguay, Chile, Panama, Colombia, and Honduras), we demonstrate how our vocabulary-discovery approach helps capture dynamic trends specific to each election. The ability to grow a vocabulary concurrently with social media trends helps capture key milestones in election campaigns.**

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 the study of how online chatter can be used to model the social, economic or political landscape of a country.

Most approaches for tracking real-world phenomena over Twitter rely on first defining a vocabulary of keywords (or hashtags) to track over the social media. This approach is typically sufficient for phenomena about which we have a fairly stable understanding. But for rapidly evolving phenomena, this approach alone is insufficient, e.g., an election season where new developments can disrupt or bolster the preferences for competing candidates among key groups or populations. In such cases, rather than using 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 we 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 as follows:

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) In contrast to traditional co-occurrence based query expansion strategies, we develop an approach that harnesses the social structure implicit in group memberships as captured through retweets, and tweet sentiment.
3) Using ten presidential elections from eight countries of Latin America (Mexico, Venezuela, Ecuador, Paraguay, Chile, Panama, Colombia and Honduras), we show how our query expansion methodology helps capture dynamic trends and improve election forecasting performance.

The rest of this paper is organized as follows. Section 2 covers the related work in query expansion, forecasting elections using social media and social group modeling. Section 3 overviews the probabilistic soft logic (PSL) framework. Section 4 provides details on how we use PSL for election vocabulary expansion. In Section 5, we present our experimental findings, using our approach on various elections. Section 6 concludes with a brief discussion.

II. Related Work

Related work can be organized into following key categories: query expansion, forecasting elections using social media, and social group modeling.

A. Query Expansion

Query expansion is a classical technique in information retrieval (IR) [5] for improving 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. Massoudi et al. [7] consider not only the co-occurrence but also the time information to score the related terms to expand the query. 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 [8]–[11], i.e., forecasting election results by assessing the popularities of candidates and their policies. Some approaches fit a regression model to opinion polls with volume of mentions and sentiment as independent variables and the opinion polls as the dependent variable [9], [11]. More sophisticated approaches [12]–[14] either model the candidates or the voters in the elections rather than compute the aggregated sentiment of the mass. Conover et al. [13] 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. [12] 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 et al. [14] 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. [15] 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. 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. Huang et al. [16], for instance, model 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 as events unfold.

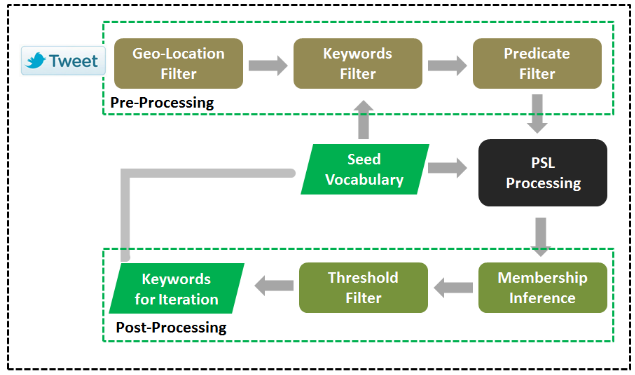


Fig. 1: Design of the query expansion pipeline.

III. Probabilistic Soft Logic

Probabilistic soft logic [4], [17] is a framework for collective probabilistic reasoning in relational domains. PSL models have been developed for various problem areas, including opinion diffusion [18], ontology alignment [17], trust in social networks [19], and knowledge graph identification [20]. 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:

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

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

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

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

where λ_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 [18], [21] 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}. 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 are used. After preliminary 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 $\text{TWEETED}(U,T)$ and $\text{CONTAINS}(T,W)$ 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 $\text{ISMEMBER}(U,G)$ and $\text{BELONGS}(W,G)$ respectively. In order to capture the temporal connectivity between the iterations, in addition to the initiating the inference process with the rule

$$\text{SEEDWORD}(W,G) \Rightarrow \text{BELONGS}(W,G)$$

we define additional rules such as

$$\text{WASMEMBER}(U,G) \Rightarrow \text{ISMEMBER}(U,G)$$

$$\text{BELONGED}(W,G) \Rightarrow \text{BELONGS}(W,G)$$

where the predicates WASMEMBER and 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.

$$\textsc{Tweeted}(U,T)\,\tilde{\wedge}\,\textsc{Contains}(T,W)\,\tilde{\wedge}\,\textsc{Belongs}(W,G)\\ \tilde{\wedge}\,\textsc{Positive}(T)\Rightarrow\textsc{IsMember}(U,G)$$

$$\textsc{Tweeted}(U,T)\,\tilde{\wedge}\,\textsc{Contains}(T,W)\,\tilde{\wedge}\,\textsc{Belongs}(W,G)\\ \tilde{\wedge}\,\textsc{Negative}(T)\Rightarrow\tilde{\neg}\,\textsc{IsMember}(U,G)$$

$$\textsc{IsMember}(U,G)\,\tilde{\wedge}\,\textsc{Tweeted}(U,T)\,\tilde{\wedge}\,\textsc{Contains}(T,W)\\ \tilde{\wedge}\,\textsc{Positive}(T)\Rightarrow\textsc{Belongs}(W,G)$$

$$\textsc{IsMember}(U,G)\,\tilde{\wedge}\,\textsc{Tweeted}(U,T)\,\tilde{\wedge}\,\textsc{Contains}(T,W)\\ \tilde{\wedge}\,\textsc{Negative}(T)\Rightarrow\tilde{\neg}\,\textsc{Belongs}(W,G)$$

Here Positive and 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 Positive and highly negative tweets are assigned a truth value of 1.0 for the predicate 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 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 Positive and 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.

$$\textsc{IsMember}(U1,G)\,\tilde{\wedge}\,\textsc{Tweeted}(U2,T)\,\tilde{\wedge}\,\textsc{ReTweet}(T,U1)\\ \Rightarrow\textsc{IsMember}(U2,G)$$

$$\textsc{IsMember}(U1,G)\,\tilde{\wedge}\,\textsc{Tweeted}(U2,T)\,\tilde{\wedge}\,\textsc{Mentions}(T,U1)\\ \tilde{\wedge}\,\textsc{Positive}(T)\Rightarrow\textsc{IsMember}(U2,G)$$

$$\textsc{IsMember}(U1,G)\,\tilde{\wedge}\,\textsc{Tweeted}(U2,T)\,\tilde{\wedge}\,\textsc{Mentions}(T,U1)\\ \tilde{\wedge}\,\textsc{Negative}(T)\Rightarrow\tilde{\neg}\,\textsc{IsMember}(U2,G)$$

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 likely about the candidate too. Since

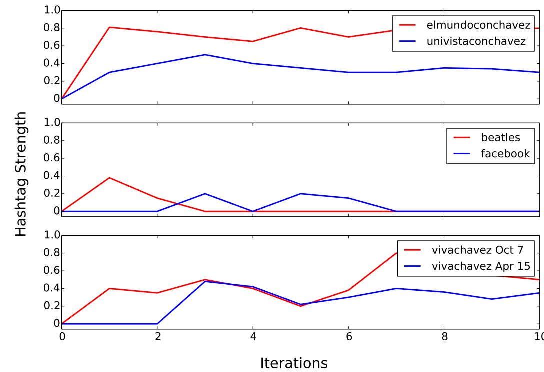


Fig. 2: Evolution of different hashtags identified for Hugo Chávez in Venezuela 2012 presidential election

these two rules have two variables W1 and W2, the number for rules generated by substituting actual groundings (hashtags) increases 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.

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

$$\textsc{Contains}(T,W1)\,\tilde{\wedge}\,\textsc{Contains}(T,W2)\,\tilde{\wedge}\\ \textsc{SeedWord}(W1,G)\,\tilde{\wedge}\,\textsc{Negative}(T)\\ \Rightarrow\tilde{\neg}\,\textsc{Belongs}(W2,G)$$

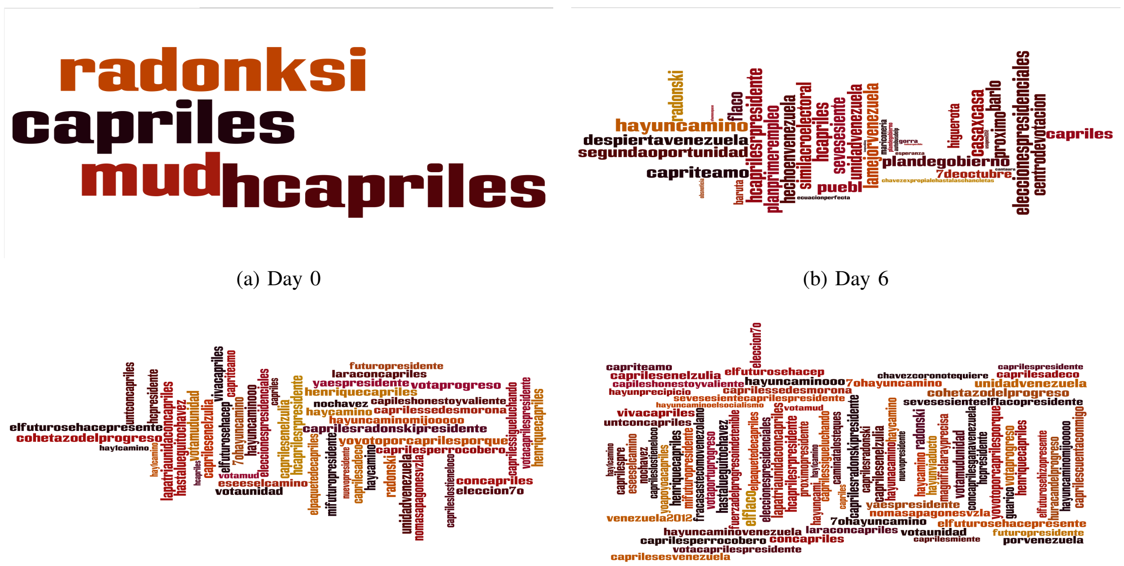
In addition to the rules, we also define constraints on the Belongs and IsMember predicates so that a particular hashtag or author can be associated to at most one candidate. Once all the tweets are loaded into the PSL program as predicates, we start the inference process by closing all the predicates 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. PSL inference then finds the most likely joint soft truth values for all the user group affiliations IsMember and the group hashtag tendencies Belongs, taking into account the various dependencies created by the weighted rules.

V. Experimental Results

Our experimental results address 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) Day 0 (b) Day 6

(c) Day 15 (d) Day 30

Fig. 3: Evolution of hashtags for Henrique Capriles in 2013 Venezuela presidential election

A. *Election vocabularies inferred*

Venezuela: Figure 3 shows how the hashtags for Henrique Capriles evolved during the month leading up to the election. Initially in Figure 3a the system begins with only a few hand picked hashtags that constitute the seed vocabulary. After a few iterations Figure 3b 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 3d). It is also interesting to note that the algorithm identified hashtags like *"nochavez"* (Figure 3c) and attributed it rightly to Hugo Chávez's primary contender, i.e., Capriles.

In Figure 2, 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 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.

Ecuador: Ecuador had a first round for the presidential election on February 17, 2013 where Rafael Correa obtained more than 50% of the votes needed to avoid a second round. The tags (Fig. 5b) include support 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).

Colombia: The hashtags identified for Juan Manuel Santos who won the Colombian Presidential elections in the second round, June 15 (first round May 25). The query expansion

(a) Nov 24

(b) Dec 15

Fig. 4: Hashtags identified for Michelle Bachelet in 2013 Chilean presidential election

(a) Mexico

(b) Ecuador

(c) Colombia

(d) Panama

Fig. 5: Vocabulary of hashtags identified for different elections

pipeline for these elections were initialized only with the main two candidates names: Santos, Manuel and Zuluaga, Ivan. We get some campain related slogans such as *"enmigobiernoprometo"* (If I govern, I promise I will). Interestingly, these kind of hashtags are often used with a sarcastic tone.

Panama: Panama's presidential election was held on May 4, with Juan Carlos Varela the winner. Seeding the query expansion pipeline were the last names of the candidates(Arias, Varela and Navarro) plus the initials of their parties(CD, PP and PRD respectively), with the addition of Arias' first name *"Domingo"* to distinguish him from a former president, Oscar Arias. From the phrase *"caminoalavictoria"* (road to victory), it indicates belief and faith that the candidate will win.

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 November 24, 2013 and the 2nd round was conducted on December 15, 2013. The query expansion pipelines 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 noise—*"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 lesser. This in turn affects the quality of the PSL inference.

B. Election Prediction

We adapt one basic forecasting algorithm from current literature to evaluate the performance of our expanded vocabularies. We emphasize that our focus in this paper is not on election forecasting per se but an interesting byproduct of our algorithm is that it provides useful features for election forecasting. To evaluate this aspect, we develop a regression approach to learn a mapping from tweet features (more below) to opinion polls, and thus forecast elections. this approach is adapted from [11] and [9].

We reason that by regressing from Twitter-derived features to the opinion polls mitigates bias caused by Twitter being a non-representative sample. We use a total of six features: Klout scores, number of unique users, total number of mentions, sentiment, and incumbency. The Klout score measures an author's influence to drive action on a scale from 1 to 100. Sentiment is scored from -20 to 20 to rate positive or negative tone associated with the tweet. Incumbency is a binary variable indicating whether the candidate is the existing holder of the political office. Moreover, the Klout scores, unique users, and mentions are further categorized into positive and negative mentions. Using our expanded vocabularies to suitably apportion tweets across the candidates, we normalize each of these features 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.

Performance: The model was tested exhaustively on a total of ten presidential elections from Latin America during 2012 and 2014. We run the prediction model in two different setting: one using the seed words and the other using PSL query expansion. As outlined in the methodology earlier, only tweets geocoded to the specific location were used to developed predictions for that location. Table I shows the overall performance of the regression model under two settings. For each election, we predict the percentage of votes gathered by each candidate. The error rate is calculated as the difference between the predicted percentage and the actual polled percentage. In all cases except the Colombia election, the PSL methodology described here gave a better estimate of the vote share, in many cases within a margin of 5%.

Election	Candidate	Actual Result	Seed Vocab.	Error	PSL Vocab.	Error
Mexico_Jul01	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_Feb17	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_Apr21	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_Nov24	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**
Panama_May04	Juan Carlos Varela	39.07	31.28	7.79	36.23	**2.84**
	Jose Domingo Arias	31.40	35.02	3.62	33.67	**2.27**
Colombia_Jun15	Juan Manuel Santos	50.95	48.85	**2.1**	45.75	5.2
	Oscar Ivan Zuluaga	45.00	43.79	**1.21**	46.72	1.72

TABLE I: Reduction in prediction error by regressing Tweet features derived from the PSL vocabulary to opinion polls. All values shown are percentages. Observe that in all but the Colombian election, the PSL vocabulary provides a better estimate of the vote share of the candidate.

VI. Discussion

We have demonstrated a novel query expansion methodology using PSL and illustrated our method could correctly capture the evolving election related vocabularies during the election season. We believe that the dynamic query expansion system is general and can be applied to not only in election but also in a lot of other domains. 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.

VII. Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. WI-IAT'10*, vol. 1. IEEE, 2010, pp. 492–499.

[2] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proc. SOMA'10*. 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. H. Bach, M. Broecheler, B. Huang, and L. Getoor, "A short introduction to probabilistic soft logic," in *Proc. NIPS'12*, 2012.

[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 *Proc. SIGIR'07*. ACM, 2007, pp. 311–318.

[7] K. Massoudi, M. Tsagkias, M. de Rijke, and W. Weerkamp, "Incorporating query expansion and quality indicators in searching microblog posts," in *AIR'11*. Springer, 2011, pp. 362–367.

[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. ICWSM'10*, vol. 10, 2010, pp. 178–185.

[9] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proc. ICWSM'10*, vol. 11, 2010, pp. 122–129.

[10] D. Saez-Trumper, W. Meira, and V. Almeida, "From total hits to unique visitors model for elections forecasting," in *Proc. ICWS'11*, 2011.

[11] A. Bermingham and A. F. Smeaton, "On using Twitter to monitor political sentiment and predict election results," in *Proc. SAAIP'11*. AFNLP, 2011.

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

[13] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in *Proc. SocialCom'11*. IEEE, 2011, pp. 192–199.

[14] E. Diaz-Aviles, C. Orellana-Rodriguez, and W. Nejdl, "Taking the pulse of political emotions in Latin America based on social web streams," in *Proc. LA-WEB'12*. IEEE, 2012, pp. 40–47.

[15] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opinions of the long tail," in *Proc. SocialCom'11*. IEEE, 2011, pp. 103–110.

[16] B. Huang, S. H. Bach, E. Norris, J. Pujara, and L. Getoor, "Social group modeling with probabilistic soft logic," in *Proc. NIPS'12*, 2012.

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

[18] S. H. 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.

[19] B. Huang, A. Kimmig, L. Getoor, and J. Golbeck, "A flexible framework for probabilistic models of social trust," in *SBP'13*, 2013.

[20] J. Pujara, H. Miao, L. Getoor, and W. Cohen, "Knowledge graph identification," in *Proc. ISWC'13*, 2013.

[21] S. H. Bach, B. Huang, B. London, and L. Getoor, "Hinge-loss Markov random fields: Convex inference for structured prediction," in *Proc. UAI'13*, 2013.