Improving Gender Prediction of Social Media Users via Weighted Annotator Rationales

Svitlana Volkova and David Yarowsky
Johns Hopkins University, Center for Language and Speech Processing
Human Language Technology Center of Excellence
Baltimore, MD 21218
svitlana@jhu.edu, yarowsky@cs.jhu.edu

Abstract

This paper proposes and contrastively evaluates several novel approaches to utilizing annotator rationales to improve the prediction of user gender in social media for English and Spanish. Our methods outperform state-of-the-art systems for Twitter gender prediction, and yield up to 28% error reduction relative to an otherwise identical system and training data without the use of annotator rationales.

1 Introduction

With the rapid growth of social media in recent years, there has been an increased interest in automatically characterizing social media users based on the informal content they generate. An important goal of this task of customer profiling or personal analytics is to label users with demographic categories, such as gender, age, ethnicity, or to determine user interests or preferences, such as political orientation, movies or product likes. Moreover, predicting user characteristics, preferences and opinions from these personalized and diverse timely data can help answer important social science questions and support many commercial applications including targeted computational advertising to match user interest profile from Twitter or Facebook,[1] detecting fraudulent product reviews [17, 13] or branding analytics [26].

There is a substantial prior work on characterizing communicants in social media, especially in Twitter. It includes inferring such latent attributes as: *gender* [19, 5, 23, 10, 4, 6], *age* [16], *political preferences* [11, 8, 18, 30, 7, 24], *personality* [12, 1, 15], *ethnicity, origin* and *race* [2].

Another promising yet understudied area of research is to elicit and utilize annotator rationales, targeted annotator feedback regarding *why/how* they chose a particular annotation. The primary example of this approach in the NLP literature is by [28], who used highlighted substrings of text as enhanced feedback to improve sentiment classification of movie reviews, with follow-on work by [29] and [27].

Given the success of Zaidan et al.'s work, and the very minimal investigation of annotator rationales in the NLP field, we have novelly applied, evaluated and substantially extended this work onto our target field of demographic prediction in social media, with additional novel contributions including:

- developing effective new ways to incorporate human domain knowledge by filtering and weighting tweets and elicited rationales in a (i) supervised and (ii) semi-supervised setting to improve user attribute classification;
- empirically assessing the benefits of the rationales and showing the advantages of rationale annotation and weighting over the state-of-the-art models for user attribute inference, in both English and Spanish.

[1] Social Network Prediction App - https://apps.facebook.com/snpredictionapp/

The cost efficiency of the proposed rationale annotation and weighting approach used in a semi-supervised bootstrapping setting will aid scaling of latent user attribute prediction to resource-limited domains and languages.

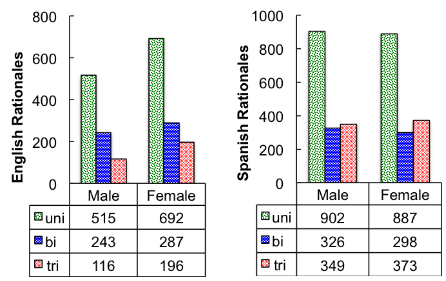
2 Data

For the experiments in this paper, we use three sets of data for each language:

I. a large pool of unlabeled data (1M tweets): for English 12.6k users with on average 78 tweets per user, and for Spanish 7.5k users with on average 132 tweets per user;
II. a small amount of training data labeled with user demographic attributes e.g., gender: for English 164 male and 193 female users, and for Spanish 251 male and 192 female users (each user is associated with 200 tweets);
III. held out test data: 100 male and 100 female users with 200 tweets per user.

The labeled training and test data is used in a supervised classification setting. The unlabeled data is used in semi-supervised setting to boost the performance of the existing supervised models for latent attribute prediction.

To collect the data we randomly sampled users from the 1% Twitter feed and downloaded 200 of their most recent tweets using the Twitter API.[2] We obtained gender labels using 3-way redundant annotation[3] on Mechanical Turk. We also asked each annotator to highlight words or phrases – *ngrams* ≤ 3 in user self-authored tweets that are highly indicative of user gender, and assign their confidence in each rationale on a 4-point scale.



(a) English (b) Spanish

Figure 1: Gender rationale ngram distribution.

Figure 2 illustrates the most frequent male and female rationales collected for English for author gender. The crowdsourced rationales resemble the results of another work that analyses language of gender in social media [21, 14, 20].[4]

We also report the distribution of rationale ngrams for both English and Spanish in Figure 1. We observe that for English the overlap of crowdsourced rationales across multiple annotators is 30.5% for unigrams, 8% for bigrams and less than 2% for trigrams. For Spanish the trend is similar.

(a) Male rationales (b) Female rationales

Figure 2: Gender rationales (word ngrams of size ≤ 3 and frequency ≥ 2). The size of each rationale reflects the frequency of being highlighted.

---

[2]Our code, data, crowdsourced annotator rationale lists for gender, age and political preference attributes as well as the detailed explanations on how we collected and annotated the data can be found here: http://www.cs.jhu.edu/ svitlana/

[3]We estimate the final label using the majority voting. The annotation agreement among three annotators exceeds 70%, and between two annotators exceeds 90%.

[4]World Well Being Project http://wwbp.org/.

3 Methodology

In this section we present our supervised and semi-supervised self-trained models with feature (rationale) weighting schemes to improve the existing approaches to author attribute classification.

3.1 Models

As input, we are given a set of users $u \in U$ represented using a multinomial distribution over user self-authored communications, e.g. their T tweets. Each user is associated with a set of 200 most recent tweets. Our goal is to predict an attribute $a \in A$ for each user $u \in U$, e.g. gender $a \in \{\text{Male, Female}\}$. For any $t \in T$, $a \in A$, the model defines a probability:

$$p(a \mid t, \vec{\theta}) = \frac{\exp(\vec{\theta} \cdot \vec{\phi}(t, a))}{\sum_{a' \in A} \exp(\vec{\theta} \cdot \vec{\phi}(t, a'))} \tag{1}$$

where $\phi : T \times A \to \mathbb{R}^d$ is a function that maps any attribute-communication pair (t, a) to a feature vector $\vec{\phi}(t, a)$. $\vec{\theta} \in \mathbb{R}^d$ is a parameter vector to learn (d is the number of features and parameters in the model); $\vec{\theta} \cdot \vec{\phi}(t, a) = \sum_{k=1}^{d} \theta_k \phi_k(t, a)$ is the inner product between $\vec{\theta}$ and $\vec{\phi}(t, a)$.

The log-linear model for such classification:

$$\Phi(u) = \begin{cases} \text{Male} & p(a \mid t, \vec{\theta}) \geq 0.5, \\ \text{Female} & \text{otherwise.} \end{cases} \tag{2}$$

Direct Model Our direct model represents a commonly-observed supervised classification setting on this task. We train our model on labeled users from TRAIN and apply it to 200 users from TEST following the Eq.1 and Eq.2. This model is learned from the labeled user tweets exclusively.

Transitive Model Given a large pool of unlabeled users and their tweets, we propose to train a direct model $\Phi(u)$ and apply it to assign labels to the thousands of unlabeled users. Then, we suggest to train a new model $\Phi_{1\text{M}}(u)$ in a semi-supervised setting, and apply both $\Phi(u)$ and $\Phi_{1\text{M}}(u)$ models to classify 200 users in TEST:

$$\Phi'(u) = \lambda \cdot \Phi(u) + (1 - \lambda) \cdot \Phi_{1\text{M}}(u) \tag{3}$$

3.2 Weighting Rationales

To incorporate attribute-specific rationales into the models defined in Eq. 1 - 3 we propose three feature weighting schemes as shown in Algorithm 1.

Algorithm 1 WEIGHTRATIONALES (r, f, ξ)

Parameters:
r: a list of rationales for each attribute value $a \in A$;
f: a list of frequencies for the rationales in r;
ξ: parameter to control rationale weights $\xi \in \{1, \ldots, 200\}$.

```
for each attribute value a ∈ A do
    for each rationale ngram r_j ∈ r do
        if (scheme == I) then
            generate ξf_j new users with r_j rationale ngrams per tweet
        else if (scheme == II) then
            generate ξ new users with f_j tweets and r_j rationale
            ngrams per tweet
        else if (scheme == III) then
            randomly sample f_j existing users for each attribute value
            a ∈ A and generate ξ tweets with r_j rationale ngrams per
            tweet
        end if
    end for
end for
```

As input we are given user self-authored communications and a list of attribute-specific rationales including m male and n female rationales $r \in R$ for gender attribute $a \in \{\text{Male}, \text{Female}\}$. The rationales r are associated with frequency $f \geq 1$. We propose to incorporate rationales into the existing models for predicting author gender using three weighting schemes described below.

For **weighting scheme I** we generate $\xi \sum_{a \in A} f \cdot r$ new data points to encode users with rationales; ξ is the parameter to be optimized. In total, we generate $\xi(m+n) \sum_{a \in A} |f|_1$ new users encoded using sparse feature vectors of ngrams. For instance, for the male rationale $r =$"gambling" with $f = 3$ and $\xi = 5$ we generate 15 new users with training instances containing the rationale ngram "gambling".

For **weighting scheme II** we generate $\xi \sum_{a \in A} r$ new data points to encode users with f rationales. In total, we generate $\xi(m+n)$ new users with less sparse feature vectors of rationales compared to the scheme A. Following the example rationale "gambling", we generate 5 new users with the training instance "gambling gambling gambling".

For **weighting scheme III** we modify f randomly sampled existing data points by adding ξ tweets with r rationales per tweet for each data point. Following the example rationale "gambling", we randomly sample 3 male users from the existing users and generate 5 training instances with the rationale ngram "gambling".

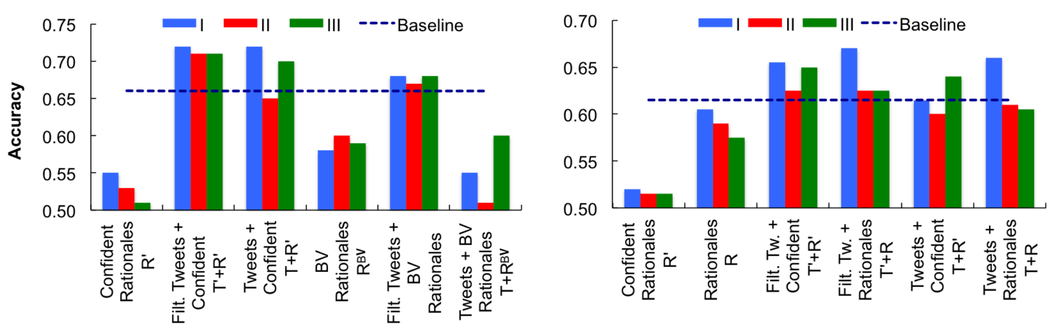
4 Experiments

4.1 Experimental Setup

We train logistic regression classifiers as shown in Eq.1 and 2 via LibLinear [9] integrated into Jerboa toolkit [22]. We optimize the classifier regularization parameters on the development data[5] and report the final results for 200 users from the test data.

4.2 Experimental Results

In Figures 3a and 3b we present accuracy results for gender classification using the baseline direct model $\Phi(u)$ defined in Eq. 2 for English and Spanish data, respectively. In contrast, we find that using only the most confident rationales (R'), with annotator confidence ≥ 3, yields lower accuracy compared to using all rationales in all other experimental variables for both languages except for some cases using weighting scheme III. Moreover, the majority our rationale weighting schemes outperform the baseline supervised model by 8% for English and 6% for Spanish in accuracy.



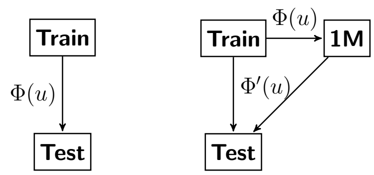
(a) English

(b) Spanish

Figure 3: Gender prediction accuracy using the direct model $\Phi(u)$ for English and Spanish.

Interestingly, we also discovered that when using rationales combined with raw tweets as user features, we could improve performance by filtering the tweets to include only those containing at least one rationale ngram $(T' + R)$ rather than using all tweets $(T + R)$. As shown in Figure 3, $(T' + R) \geq (T + R) \geq T \geq R$. The trend is the same for using only confident rationales

[5] We randomly sample 20% of the training data as development data. The remaining disjoint 80% is used for training.



	I	II	III	I	II	III
T	0.66	0.66	0.66	0.65	0.65	0.65
R	0.61	0.55	**0.63**	**0.65**	0.55	0.56
$T+R$	**0.72**	0.71	0.69	**0.74**	0.72	0.69
$T'+R$	**0.74**	0.71	0.69	0.74	0.70	**0.75**
ΔE_{max}	24%	15%	9%	26%	20%	28%

Table 1: Gender classification results for English using $\Phi(u)$, $\Phi'(u)$ models with weighted annotator rationales. Models are trained on T: tweets only, R: rationales only, $T + R$: all tweets + rationales, $T' + R$: filtered tweets + rationales. ΔE_{max} is a relative error reduction of $T' + R$ or $T + R$ compared to T.

R'. For example, $\Phi(u)$ model trained for English using weighting scheme I yields the results: $0.74 > 0.72 > 0.66 > 0.61$. Similarly, $\Phi(u)$ trained for Spanish using weighting scheme I yields the results: $0.67 > 0.65 > 0.61 > 0.60$. We get these improvements because (a) more data is better and (b) features are less sparse and highly discriminative features e.g., rationales are ranked higher compared to all other features.

In addition, we made a comparison with an contrastive distilled-feature resource, the list of conceptual class attributes for gender collected by [4]. The list contains 958 Male and 659 Female ngrams. In Figure 3 we refer to them as R^{BV} rationales. We finds that R^{BV} features perform significantly better than confident rationales but significantly worse (schema I) or comparably (schemes II and III) to using all rationales when models are learned from rationales only. When we combine tweets with rationales, models learned from our rationale plus a tweet mix $T + R$ and $T' + R$ significantly outperform the models learned from the tweet plus R^{BV} mix for all weighting schemes.

Finally, we report experimental results for English $\Phi(u)'$ models in Table 1. We find that $\Phi(u)'$ models trained in semi-supervised setting exclusively on tweets T do not yield statistically significant improvements over the baseline $\Phi(u)$. However, when user tweets are combined with rationales $T + R$ the absolute gain is 2% when weighting scheme I is applied. Moreover, when the tweets filtered to only those tweets that contain rationale ngrams (T') are mixed with raw rationales to train the model $\Phi(u)'$ with weighting scheme III, the absolute gain is the highest – 10% over the baseline T (the error reduction is $\Delta E = 28\%$).

5 Related Work

The majority of the existing models for latent author attribute or personalized preference prediction e.g., gender, age or political affiliation define this task as a supervised classification. They rely on thousands of user self-authored tweets (primarily in English) trained using bag-of-word (BoW) lexical features. For example, [23], [10], [16], and [7] rely on word ngrams. Limited works apply network structure [18], communication behavior, socio-linguistic [19], syntactic and stylistic features [3], or study gender prediction for languages other than English [6, 25]. For example, [6] report comparable to our classification accuracy for Spanish – 0.76 for French and 0.63 for Japanese.

Approach	Users	Tweets	Features	Accuracy
Rao et al. [19]	1K	0.4M (4K)	BoW, socio-ling, combined	**0.69** 0.71 0.72
Burger et al. [5]	184K	4M (22)	user names, char ngrams	0.92 **0.75**
Bergsma et al. [4]	400 1M	4B (500)	bootstrapped bootstacked	**0.72** 0.87
This work	357	70K (200)	T (only) $T' + R$	**0.66** **0.75**

Table 2: The overview of the existing approaches for gender classification in social media.

To compare our models with the existing approaches for gender prediction on Twitter we present a brief quantitative comparison in Table 2. These models are all trained in a supervised setting with various feature combinations, with the comparable bag-of-words feature performances marked in bold. Our best model outperforms the BoW baseline presented by [19] by an absolute 6%, as well as their other feature combinations by 3%. Moreover, we achieve comparable accuracy with the similar character ngram model presented by [5], but learned from millions of tweets for 184K users.

Furthermore, our work achieves 3% absolute performance gain relative to the bootstrapped models presented by [4] using conceptual class attributes over the same amount of training examples (400 users). Only when their models are bootstrapped from billions of tweets does the final accuracy increases to 0.87; we assume dramatically less annotated data.

6 Conclusions

We proposed several readily-replicable new models for gender classification of social media users for English and Spanish that outperform the state-of-the-art models learned exclusively from user data. We introduced three novel rationale weighting schemes integrated into different models with varied amount of supervision. We found that:

- T vs. $T' + R$: incorporating rationales as additional informative features into the models is beneficial for gender prediction either in fully supervised (the largest relative error rate reduction is 24%) or semi-supervised bootstrapping setting (the largest relative error rate reduction is 28%);
- R' vs R: using all rationales is better than using just confident rationales: 2 - 12% accuracy gain for English and 6 - 9 % for Spanish;
- T vs. R: in the common experimental setting where the collected Twitter data cannot be shared with others, distilled rationales alone can be used to train the models leading to only a 3% absolute accuracy loss for English and 1% for Spanish.
- $T + R$ vs. $T' + R$: applying rationales in combination with filtered tweets is better than mixing rationales with all tweets available for a given user – up to 3% absolute accuracy gain for English and 1.5% for Spanish.

Finally, the investment in rationale annotation is very cost-effective; a 28% relative error reduction is achieved with only a $10 total additional Mechanical Turk cost to collect the rationales in this reported experimental setup. Furthermore, the value of using rationales to improve performance on this task is not only about money; many domains have limited raw data or severely volume limited APIs or IP constraints, making our demonstrated rationale-based performance gains with no additional raw data even more valuable.

References

[1] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 24–32, 2012.

[2] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1010–1019, 2013.

[3] Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 327–337, 2012.

[4] Shane Bergsma and Benjamin Van Durme. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 710–720, 2013.

[5] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1301–1309, 2011.

[6] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, 2013.

[7] Raviv Cohen and Derek Ruths. Classifying Political Orientation on Twitter: It's Not Easy! In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 91–99, 2013.

[8] Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of Twitter users. In *Proceedings of Social Computing*, pages 192–199, 2011.

[9] Rong En Fan, Kai Wei Chang, Cho Jui Hsieh, Xiang Rui Wang, and Chih Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[10] Katja Filippova. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1478–1488, 2012.

[11] Jennifer Golbeck and Derek Hansen. Computing political preference among Twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1108, 2011.

[12] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from Twitter. In *Proceedings of SocialCom/PASSAT*, 2011.

[13] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining (WSDM)*, pages 219–230, 2008.

[14] Margaret Kern, Johannes Eichstaedt, Andrew Schwartz, Lukasz Dziurzynski, Lyle Ungar, David Stillwell, Michal Kosinski, Stephanie Ramones, and Martin Seligman. The online social self an open vocabulary approach to personality. *Assessment*, 21(2):158–169, 2014.

[15] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 2013.

[16] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 439–448, 2013.

[17] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 309–319, 2011.

[18] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 430–438, 2011.

[19] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents (SMUC)*, pages 37–44, 2010.

[20] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Hansen Andrew Schwartz. Developing age and gender predictive lexica over social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[21] Hansen Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[22] Benjamin Van Durme. Jerboa: A toolkit for randomized and streaming algorithms. Technical report, HLTCOE, 2012.

[23] Benjamin Van Durme. Streaming analysis of discourse participants. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 48–58, 2012.

[24] Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–196, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[25] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, 2013.

[26] William Yang Wang, Edward Lin, and John Kominek. This text has the scent of starbucks: A laplacian structured sparsity model for computational branding analytics. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, WA, USA, October 2013. ACL.

[27] Ainur Yessenalina, Yejin Choi, and Claire Cardie. Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 336–341, 2010.

[28] Omar Zaidan, Jason Eisner, and Christine D. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *HLT-NAACL*, pages 260–267, 2007.

[29] Omar F. Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 31–40, 2008.

[30] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 387–390, 2012.