Weakly Supervised Tweet Stance Classification by Relational Bootstrapping

Javid Ebrahimi and **Dejing Dou** and **Daniel Lowd**
Department of Computer and Information Science, University of Oregon
Eugene, Oregon 97403, USA
{javid,dou,lowd}@cs.uoregon.edu

Abstract

Supervised stance classification, in such domains as Congressional debates and online forums, has been a topic of interest in the past decade. Approaches have evolved from text classification to structured output prediction, including collective classification and sequence labeling. In this work, we investigate collective classification of stances on Twitter, using hinge-loss Markov random fields (HL-MRFs). Given the graph of all posts, users, and their relationships, we constrain the predicted post labels and latent user labels to correspond with the network structure. We focus on a weakly supervised setting, in which only a small set of hashtags or phrases is labeled. Using our relational approach, we are able to go beyond the stance-indicative patterns and harvest more stance-indicative tweets, which can also be used to train any linear text classifier when the network structure is not available or is costly.

1 Introduction

Stance classification is the task of determining from text whether the author of the text is in favor of, against, or neutral towards a target of interest. This is an interesting task to study on social networks due to the abundance of personalized and opinionated language. Studying stance classification can be beneficial in identifying electoral issues and understanding how public stance is shaped (Mohammad et al., 2015).

Twitter provides a wealth of information: public tweets by individuals, their profile information, whom they follow, and more. Exploiting all these pieces of information, in addition to the text, could help build better NLP systems. Examples of this approach include user preference modeling (Li et al., 2014), stance classification (Rajadesingan and Liu, 2014), and geolocation identification (Jurgens, 2013; Rahimi et al., 2015). For stance classification, knowing the author's past posting behavior, or her friends' stances on issues, could improve the stance classifier. These are inherently structured problems, and they demand structured solutions, such as Statistical Relational Learning (SRL) (Getoor, 2007). In this paper, we use hinge-loss Markov random fields (HL-MRFs) (Bach et al., 2015), a recent development in the SRL community.

SemEval 2016 Task 6 organizers (Mohammad et al., 2016) released a dataset with Donald Trump as the target, without stance annotation. The goal of the task was to evaluate stance classification systems, which used minimal labeling on phrases. This scenario is becoming more and more relevant due to the vast amount of data and ever-changing nature of the language on social media. This is critical in applications in which a timely detection is highly desired, such as violence detection (Cano Basave et al., 2013) and disaster situations.

Our work is the first to use SRL for stance classification on Twitter. We formulate the weakly supervised stance classification problem as a bi-type collective classification problem: We start from a small set of stance-indicative patterns and label the tweets as positive and negative, accordingly. Then, our relational learner uses these noisy-labeled tweets, as well as the network structure, to classify the stance

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1012–1017,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

of other tweets and authors. Our goal will be to constrain pairs of similar tweets, pairs of tweets and their authors, and pairs of neighboring users to have similar labels. We do this through hinge-loss feature functions that encode our background knowledge about the domain: (1) A person is pro/against Trump if she writes a tweet with such stance; (2) Friends in a social network often agree on their stance toward Trump; (3) similar tweets express similar stances.

2 Related Work

Stance classification is related to sentiment classification with a major difference that the target of interest may not be explicitly mentioned in the text and it may not be the target of opinion in the text (Mohammad et al., 2016). Previous work has focused on Congressional debates (Thomas et al., 2006; Yessenalina et al., 2010), company-internal discussions (Agrawal et al., 2003), and debates in online forums (Anand et al., 2011; Somasundaran and Wiebe, 2010). Stance classification has newly been posed as structured output prediction. For example, citation structure (Burfoot et al., 2011) or rebuttal links (Walker et al., 2012) are used as extra information to model agreements or disagreements in debate posts and to infer their labels. Arguments and counterarguments occur in sequences; Hasan and Ng (2014) used this observation and posed stance classification in debate forums as a sequence labeling task, and used a global inference method to classify the posts.

Sridhar et al. (2015) use HL-MRFs to collectively classify stance in online debate forums. We address a weakly supervised problem, which makes our approach different as we do not rely on local text classifiers. Rajadesingan et al. (2014) propose a retweet-based label propagation method which starts from a set of known opinionated users and labels the tweets posted by the people who were in the retweet network.

3 Stance Classification on Twitter

3.1 Markov Random Fields

Markov random fields (MRFs) are widely used in machine learning and statistics. Discriminative Markov random fields such as conditional random fields (Lafferty et al., 2001) are defined by a joint distribution over random variables $Y_1, ..., Y_m$ conditioned on $X_1, ..., X_n$ that is specified by a vector of d real-valued potential functions $\phi_l(\boldsymbol{y}, \boldsymbol{x})$ for $l = 1, ..., d$, and a parameter (weight) vector $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{x})}\exp\Big(\langle \boldsymbol{\theta}, \phi(\boldsymbol{y}, \boldsymbol{x})\rangle\Big)$$

where $\langle \boldsymbol{\theta}, \phi(\boldsymbol{y}, \boldsymbol{x})\rangle$ denotes the dot product of the parameters and the potential functions, and $Z(\boldsymbol{\theta}, \boldsymbol{x})$ is the partition function.

3.2 HL-MRFs for Tweet Stance Classification

Finding the *maximum a posteriori* (MAP) state is a difficult discrete optimization problem and, in general, is NP-hard. One particular class of MRFs that allows for convex inference is hinge-loss Markov random fields (HL-MRFs) (Bach et al., 2015). In this graphical model, each potential function is a hinge-loss function, and instead of discrete variables, MAP inference is performed over relaxed continuous variables with domain $[0, 1]^n$. These hinge-loss functions, multiplied by the corresponding model parameters (weights), act as penalizers for soft linear constraints in the graphical model.

Consider $\boldsymbol{t}_i$, $\boldsymbol{u}_j$ as the random variables denoting the ith tweet and the jth user. The potential function, $\phi(\boldsymbol{t}_i, \boldsymbol{u}_j)$, relating a user and her tweet is as follows,

$$\max(0, t_{ik} - u_{jk}) \tag{1}$$

where t_{ik} and u_{jk} denote the respective assertions that $\boldsymbol{t}_i$ has label k, and $\boldsymbol{u}_j$ has label k . The function captures the distance between the label for a user and her tweet. In other words, this function measures the penalty for dissimilar labels for a user and her tweet.

For users who are “friends” (i.e., who “follow” each other on Twitter), we add this potential function,

$$\max(0, u_{ik} - u_{jk}) \tag{2}$$

and for the tweet-tweet relations,

$$s_{ij}\max(0, t_{ik} - t_{jk}) \tag{3}$$

where s_{ij} measures the similarity between two tweets. This scalar helps penalize violations in proportion to the similarity between the tweets. For the similarity measure, we simply used the cosine similarity between the n-gram (1-4-gram) representation of the tweets and set 0.7 as the cutoff threshold.

Finally, two hard linear constraints are added, to ensure that $\boldsymbol{t}_i$, and $\boldsymbol{u}_j$ are each assigned a single label, or in other words, are fractionally assigned labels with weights that sum to one.

$$\sum_k t_{ik} = 1 \;,\; \sum_k u_{ik} = 1 \tag{4}$$

Weight learning is performed by an improved structured voted perceptron (Lowd and Domingos, 2007), at every iteration of which we estimate the labels of the users by hard EM. This formulation can work in weakly supervised settings, because the constraints simply dictate similar/neighboring nodes to have similar labels.

In the language of Probabilistic Soft Logic (PSL) (Bach et al., 2015), the constraints can be defined by the following rules:

PSL Rules:

tweet-label(T, L) $\wedge$ tweet-user(T, U) $\Rightarrow$ user-label(U, L)
user-label(U_1, L) $\wedge$ friend(U_1, U_2) $\Rightarrow$ user-label(U_2, L)
tweet-label(T_1, L) $\wedge$ similar(T_1, T_2) $\Rightarrow$ tweet-label(T_2, L)
PredicateConstraint.Functional , on : user-label
PredicateConstraint.Functional , on : tweet-label

Our post-similarity constraint implementation is different from the original PSL implementation due to the *multiplicative* similarity scalar[1].

This work is a first step toward relational stance classification on Twitter. Incorporating other relational features, such as mention networks and retweet networks can potentially improve our results. Similarly, employing textual entailment techniques for tweet similarity will most probably improve our results.

4 Experiments and Results

4.1 Data

SemEval-2016 Task 6.b (Mohammad et al., 2016) provided 78,000+ tweets associated with "Donald Trump". The protocol of the task only allowed minimal manual labeling, i.e. "tweets or sentences that are manually labeled for stance" were not allowed, but "manually labeling a handful of hashtags" was permitted. Additionally, using Twitter's API, we collected each user's follower list and their profile information. This often requires a few queries per user. We only considered the tweets which contain no URL, are not retweets, and have at most three hashtags and three mentions.

Algorithm Relational Bootstrapping

```
Input:
Unlabeled pairs of tweets and authors ($\mathbf{t}_i$, $\mathbf{u}_i$).
Friendship pairs ($\mathbf{u}_i$, $\mathbf{u}_j$) between users.
Similarity triplets ($\mathbf{t}_i$, $\mathbf{t}_j$, $s_{ij}$) between tweets.
Stance-indicative regexes $\mathbf{R}$.
// Create an initial dataset.
Training set $\mathbf{X}$ = $\{\}$.
Harvest positive and negative tweets based on $\mathbf{R}$.
Add the harvested tweets to $\mathbf{X}$.
// Augment the dataset by the relational classifier.
Learning & inference over $P(\mathbf{U}, \mathbf{T}|\mathbf{X})$ by our HL-MRF.
Add some classified tweets to training set: $\mathbf{X} = \mathbf{X} + \mathbf{T}$.
Output: $\mathbf{X}$.
```

Favor.	make(?)america(?)great(?)again, #trumpforpresident, I{'m, am} voting trump, #illegal(.*), patriot, #boycottmacy
Against.	racist, bigot, idiot, hair, narcissis(.+)

Table 1: Patterns to collect pro-Trump and anti-Trump tweets.

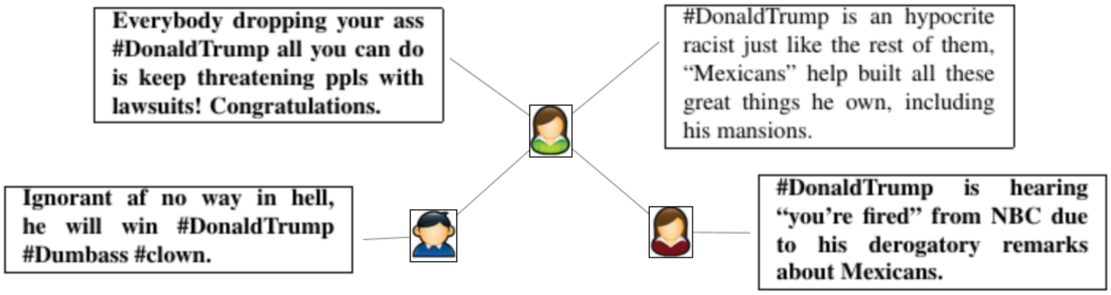
This task's goal was to test stance towards the target in 707 tweets. The authors in the test set are not identified, which prevents us from pursuing a fully relational approach. Thus, we adopt a two-phase approach: First, we predict the stance of the training tweets using our HL-MRF. Second, we use the labeled instances as training for a linear text classifier. This dataset-augmenting procedure is summarized in the Algorithm *Relational Bootstrapping*.

4.2 Experimental Setup

We pick the pro-Trump and anti-Trump indicative regular expressions and hashtags, which are shown in Table 1. Tweets that have at least one positive or one negative pattern, and do not have both positive and negative patterns, are considered as our initial positive and negative instances. This gives us a dataset with noisy labels; for example, the tweet "*his #MakeAmericaGreatAgain #Tag is a bummer.*" is against Donald Trump, incorrectly labeled favorable. A quantitative analysis of the impact of noise, and the goodness of initial patterns, can be pursued in the future through a supervised approach.

Tweets in the "neither" class range from news about the target of interest, to tweets totally irrele-

[1] The original implementation would result in the function, max(0, $t_{ik} + s_{ij} - t_{jk} - 1$), which is less intuitive than ours.


Everybody dropping your ass #DonaldTrump all you can do is keep threatening ppls with lawsuits! Congratulations.

#DonaldTrump is an hypocrite racist just like the rest of them, "Mexicans" help built all these great things he own, including his mansions.

Ignorant af no way in hell, he will win #DonaldTrump #Dumbass #clown.

#DonaldTrump is hearing "you're fired" from NBC due to his derogatory remarks about Mexicans.

Figure 1: An example of the output of our relational bootstrapper. A small excerpt of the network, consisting of three users, four tweets and two friendship links. The tweet in regular type face is labeled as anti-Trump in the first phase, because of the word "racist" in the tweet. The other tweets, which are in boldface, are found through SRL harvesting, and are automatically labeled as anti-Trump tweets correctly.

vant to him. This makes it difficult to collect neutral tweets, and we will classify tweets to be in that class based on a heuristic described in the next subsection.

Given the limited number of seeds, we need to collect more training instances to build a stance classifier. Because of the original noise in the labels and the imposed fragmentary view of data, *self-learning* would perform poorly. Instead, we augment the dataset with tweets that our relational model classifies as positive or negative with a minimum confidence (class value 0.52 for pro-Trump and 0.56 for anti-Trump). The hyper-parameters were found through experimenting on a development set, which was the stance-annotated dataset of SemEval Task 6.a. The targets of that dataset include Hillary Clinton, Abortion, Climate Change, and Athesim. Since there are more anti-Trump tweets than pro-Trump (Mohammad et al., 2016), for our grid search we prefer a higher confidence threshold for the anti-Trump class, making it harder for the class bias to adversely impact the quality of harvested tweets. We also exclude the tweets that were sent by a user with no friends in the network. An example which showcases relational harvesting of tweets can be seen in Figure 1, wherein given the evidence, some of which is shown, three new tweets are found.

No. total tweets	21,000
No. initial pro tweets	1,100
No. initial anti tweets	1,490
No. relational-harvested pro tweets	960
No. relational-harvested anti tweets	780
No. edges in tweet similarity network	7,400
No. edges in friend network	131,000

Table 2: Statistics of the data

4.3 Classification

We convert the tweets to lowercase, and we remove stopwords and punctuation marks. For tweet classification, we use a linear-kernel SVM, which has proven to be effective for text classification and robust in high-dimensional spaces. We use the implementation of Pedregosa et al. (2011), and we employ the features below, which are normalized to unit length after conjoinment.

N-grams: `tf-idf` of binary representation of word n-grams (1–4 gram) and character n-grams (2–6 gram). After normalization, we only pick the top 5% most frequent grams.

Lexicon: Binary indicators of positive-emotion and negative-emotion words in LIWC2007 categories (Tausczik and Pennebaker, 2010).

Sentiment: Sentiment distribution, based on a sentiment analyzer for tweets, VADER (Hutto and Gilbert, 2014).

Table 3 demonstrates the results of stance classification. The metrics used are the macro-average of the F1-score for favor, against, and average of these two. The best competing system for the task used a deep convolutional neural network to train on pro and against instances, which were collected through linguistic patterns. At test time, they randomly assigned the instances, about which the classifier was less confident, to the "neither" class. Another base-

Method	F_{favor}	$F_{against}$	F_{avg}
SVM-ngrams-comb	18.42	38.45	28.43
best-system	**57.39**	55.17	56.28
SVM-IN	30.43	59.52	44.97
SVM-NB	47.67	57.53	52.60
SVM-RB	52.14	59.26	55.70
SVM-RB-N	54.27	**60.77**	**57.52**

Table 3: Evaluation on SemEval-2016 Task 6.b.

line is an SVM, trained on another stance classification dataset (Task 6.a), using a combination of n-gram features (SVM-ngrams-comb).

SVM-IN is trained on the initial dataset created by linguistic patterns, SVM-RB is trained on the relational-augmented dataset, and SVM-NB is a naive bootstrapping method that simply adds more instances, from the users in the initial dataset, with the same label as their tweets in the initial dataset, and for those who have both positive and negative tweets, does not add more of their tweets.

At test time, we could predict an instance to be of the "neither" class if it contains none of our stance-indicative patterns, nor any of the top 100 word grams that have the highest `tf-idf` weight in the training set. SVM-RB-N follows this heuristic for the "neither" class, while SVM-RB ignores this class altogether.

4.4 Demographics of the Users

As an application of stance classification, we analyze the demographics of the users based on their profile information. Due to the demographics of Twitter users, one has to be cautious about drawing generalizing conclusions from the analysis of Twitter data. We pick a balanced set of 1000 users with the highest degree of membership to any of the two groups. In Figure 2, we plot states represented by at least 50 users in the dataset. We can see that the figure correlates with US presidential electoral politics; supporters of Trump dominate Texas, and they are in the clear minority in California.

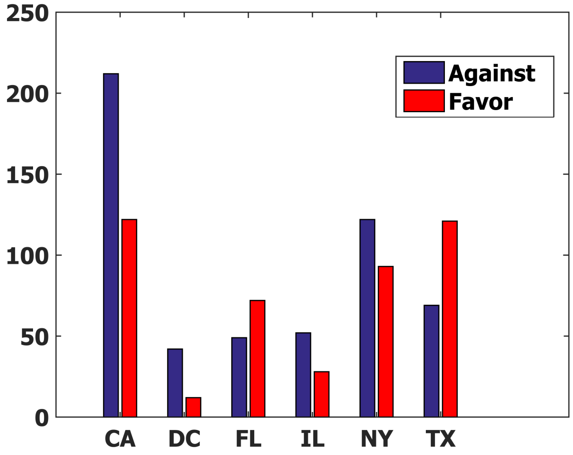


Figure 2: Distribution of Twitter users in a number of states.

5 Conclusions and Future Work

In this paper, we propose a weakly supervised stance classifier that leverages the power of relational learning to incorporate extra features that are generally present on Twitter and other social media, i.e., authorship and friendship information. HL-MRFs enables us to use a set of hard and soft linear constraints to employ both the noisy-labeled instances and background knowledge in the form of soft constraints for stance classification on Twitter.

While the relational learner tends to smooth out the incorrectly labeled instances, this model still suffers from noise in the labels. Labeling features and enforcing model expectation can be used to alleviate the impact of noise; currently, the initial linguistic patterns act as hard constraints for the label of the tweets, which can be relaxed by techniques such as generalized expectation (Druck et al., 2008).

The SemEval dataset has only one target of interest, Donald Trump. But the target of the opinion in the tweet may not necessarily be him, but related targets, such as Hillary Clinton and Ted Cruz. Thus, automatic detection of targets and inferring the stance towards all of the targets is the next step toward creating a practical weakly-supervised stance classifier.

6 Acknowledgments

This work was supported by NIH grant R01GM103309 and ARO grant W911NF-15-1-0265. We would like to thank anonymous reviewers for their helpful comments, Saed Rezayi for helping with Twitter API, and Ellen Klowden for discussions.

References

Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, and Yirong Xu. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of WWW*, pages 529–535.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 1–9.

Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2015. Hinge-loss Markov random fields and probabilistic soft logic. arXiv:1505.04406 [cs.LG].

Clinton Burfoot, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of ACL*, pages 1506–1515.

Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Proceedings of IJCNLP*, pages 109–117.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of SIGIR*, pages 595–602.

Lise Getoor. 2007. *Introduction to statistical relational learning*. MIT press.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP*, pages 751–762.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of ICWSM*, pages 216–225.

David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of ICWSM*, pages 273–282.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

Jiwei Li, Alan Ritter, and Dan Jurafsky. 2014. Inferring user preferences by probabilistic logical reasoning over social networks. arXiv:1411.2679 [cs.SI].

Daniel Lowd and Pedro Domingos. 2007. Efficient weight learning for Markov logic networks. In *Proceedings of PKDD*, pages 200–211.

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval*, pages 31–41.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of ACL*, pages 630–636.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *Proceedings of SBP*, pages 153–160.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*, pages 116–125.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of NAACL-HLT*, pages 592–596.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of EMNLP*, pages 1046–1056.