A Joint Sentiment-Target-Stance Model for Stance Classification in Tweets

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

Abstract

Classifying the stance expressed in online microblogging social media is an emerging problem in opinion mining. We propose a probabilistic approach to stance classification in tweets, which models stance, target of stance, and sentiment of tweet, jointly. Instead of simply conjoining the sentiment or target variables as extra variables to the feature space, we use a novel formulation to incorporate three-way interactions among sentiment-stance-input variables and three-way interactions among target-stance-input variables. The proposed specification intuitively aims to discriminate sentiment features from target features for stance classification. In addition, regularizing a single stance classifier, which handles all targets, acts as a soft weight-sharing among them. We demonstrate that discriminative training of this model achieves the state-of-the-art results in supervised stance classification, and its generative training obtains competitive results in the weakly supervised setting.

1 Introduction

Stance Classification (SC) is the task of inferring from text whether the author is in favor of a given target, against it, or has a neutral position toward it. This task, which can be complex even for humans (Walker et al., 2012a), is related to argument mining, subjectivity analysis, and sentiment classification. Generic sentiment classification is formulated as determining whether a piece of text is positive, negative, or neutral. However, in SC, systems must detect favorability toward a given (pre-chosen) target of interest. In this sense, SC is more similar to target-dependent sentiment classification (Jiang et al., 2011), with a major difference that the target of the stance might not be explicitly mentioned in text or might not be the target of the opinion (Mohammad et al., 2016). For example, the tweet below implies a stance against Donald Trump, through expressing support for Hillary Clinton.

> Target: Donald Trump
> My vote is definitely for Hillary. Can't trust #gop candidates.

This is an interesting task to study on social networks because of the abundance of personalized and opinionated language. Given the growing significance of the role social media is playing in our world, studying stance classification can be beneficial among others, in identifying electoral issues and understanding how public stance is shaped (Mohammad et al., 2015).

SemEval 2016 Task 6 organizers (Mohammad et al., 2016) released a joint stance and sentiment annotated dataset. Studying the correlation between sentiment and stance and how the former can help detect the latter is an important research question that we address in this paper. Our approach relies on one observation for stance detection in tweets. Ignoring general words and stopwords, a lot of the time, we can expect a rough dichotomy on the remaining n-grams of the tweets. Concretely, a stance-related n-gram either refers to a topic related to the target or bears a sentiment. In Table 1 *Christian*, *religion*, *Feminism*, and *campaign* are of the first type, while *murder* and *enjoyed* are of the second type. We design the model such that the probability of a stance y given the text x, and its associated target t and

(1) **Target**: legalization of abortion, **Stance**:Against, **Sentiment**:Negative Hillary, Here's one Christian whose religious views will never adapt to include abortion. Abortion is murder. (2) **Target**: Hillary Clinton, **Stance**:Favor, **Sentiment**:Positive Enjoyed @jamiaw article on feminism + @hillaryclinton. We are building campaign that engages ppl through an intersectional lens.

Table 1: Two examples from SemEval 2016 Task 6.A on two of the targets specified in the corpus.

sentiment s, is proportional to the product of two components. The first component measures the consistency of x with sentiment s and stance y, while the second component measures the consistency of x with target t and stance y. The model learns how to discriminate among the target-specific features and sentiment-specific features, the latter of which might be generalized across different targets. This is further improved by performing regularization on one single classifier, as opposed to a different classifier for each target, which so far has been the standard way to do stance classification.

Our discriminative model works effectively for supervised stance classification tasks. However, manual annotation requires painstaking work by researchers, which can be even more difficult for tasks such as sentiment annotation (Mohammad, 2016). To this end, we propose a generative model, which works properly for stance prediction especially in weakly supervised settings, in which labeled instances are few and labels might be noisy.

Our contributions are as follows:

1. We address the modeling of interactions among target of stance, stance itself, and sentiment in text, by an undirected graphical model.

2. We use one single classifier for stance classification across multiple targets, as opposed to previous works, which use a separate classifier for each target. We demonstrate how our particular model specification and shared regularization can improve stance classification across multiple targets.

3. We develop both discriminative and generative training algorithms, which achieve the state-of-the-art results on supervised and competitive results for weakly supervised stance classification tasks, respectively.

2 Related Work

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). There is a growing interest in performing stance classification on other media. For example, Faulkner (2014) detected document-level stance in student essays. Sobhani et al. (2015) extracted arguments used in online news comments to detect stance. The data from the Emergent Project[1] was used to classify the stance of article headlines (Ferreira and Vlachos, 2016). SemEval-2016 Task 6 (Mohammad et al., 2016) involved two stance detection subtasks in tweets in supervised and weakly supervised settings.

Somasundaran and Wiebe (2010) developed a baseline for stance classification using features based on modal verbs and sentiments. Anand et al. (2011) augmented the *n*-gram features with lexicon-based and dependency-based features. FrameNet semantic frames have also been incorporated in (Hasan and Ng, 2013; Hasan and Ng, 2014). SC has newly been posed as collective classification. For example, citation structure (Burfoot et al., 2011) or rebuttal links (Walker et al., 2012b), are used as extra information to model agreements or disagreements in debate posts and to infer their labels. In (Murakami and Raymond, 2010) a *maximum cut* method is used to aggregate stances in multiple posts to infer a user's stance on the target. Sridhar et al. (2015) use Probabilistic Soft Logic (PSL) to collectively classify the stance of users and stance in posts. PSL has also been used to augment a weakly-labeled tweet collection by incorporating Twitter's network-based features (Ebrahimi et al., 2016). Similarly, Rajadesingan et al.

[1] http://towcenter.org/research/lies-damn-lies-and-viral-content/

(2014), use a retweet-based label propagation method, which starts from a set of opinionated users and labeled tweets by the people who are in the retweet network. Since arguments and counter-arguments occur in sequences, Hasan and Ng (2014) were able to pose stance classification in debate forums as a sequence labeling task.

Tweets pose some challenges that preclude the use of standard off-the-shelf NLP feature extractors (Dey and Haque, 2009). Tweets have restricted length, which sometime makes the author use unstructured or incoherent statements. This is aggravated by the highly informal language that is common on Twitter, which includes grammatical errors. Not surprisingly, the results of SemEval 2016 Task 6 (Mohammad et al., 2016) showed the effectiveness of simple word *n*-grams and character *n*-grams, in addition to deep neural network approaches that automatically extract features. We use *n*-gram features in this work. But we will briefly discuss how our (discriminative) formulation can be incorporated into neural nets too.

In the next sections, we present effective baselines for joint modeling of targets, sentiments, and stances by a simple log-linear approach. We develop both generative and discriminative models and perform experiments on SemEval 2016 Tasks 6.A and 6.B (Mohammad et al., 2016).

3 STS: Joint Sentiment-Target-Stance Modeling

3.1 Log-Linear STS Model

In this paper, $\mathbf{x} \in \mathbb{R}^V$ is a vector of input features and $y \in Y$ is a discrete stance label, which we handle by a one-hot vector $\boldsymbol{e}_y \in \mathbb{R}^M$. Similar vectors, $\boldsymbol{e}_s \in \mathbb{R}^P$ and $\boldsymbol{e}_t \in \mathbb{R}^Q$, are defined for sentiment and target variables respectively. The model is defined over tensors $\Lambda^1 \in \mathbb{R}^{P \times V \times M}$ and $\Lambda^2 \in \mathbb{R}^{Q \times V \times M}$. Λ^1 and Λ^2 govern the sentiment-stance-feature and target-stance-feature interactions respectively. We define the following energy function,

$$E(y|\mathbf{x}, s, t; \Lambda^1, \Lambda^2) = -\sum_k \mathbf{x}_k(\lambda^1_{s,k,y} + \lambda^2_{t,k,y})$$

The negative of the energy associated with the stance label y, given the text $\mathbf{x}$, and its associated target t and sentiment s, is equal to the summation of two components. The first one measures the consistency of $\mathbf{x}$ with sentiment s and stance y, while the second one measures the consistency of $\mathbf{x}$ with target t and stance y. This specification is a log-linear model with feature functions $\phi^1(t, k, y) = \mathbb{1}(T = t, \mathbf{x}_k = 1, Y = y)$ and $\phi^2(s, k, y) = \mathbb{1}(S = s, \mathbf{x}_k = 1, Y = y)$, where $\mathbb{1}$ is the indicator function.

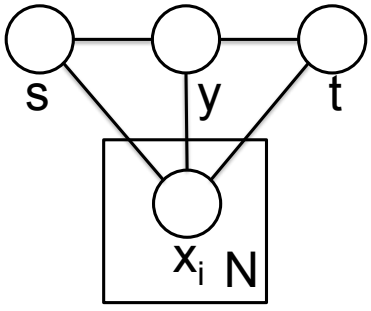

Figure 1: Plate model for STS

We build $p_\theta(y, s, t, \mathbf{x})$, a generatively trained model, and $p_\theta(y|s, t, \mathbf{x})$, a discriminatively trained model. For both models, the inference problem, the probability of output y conditioned on inputs, is given by,

$$p(y|\mathbf{x}, s, t) = \frac{\exp\big(\mathbf{x}^T(\boldsymbol{e}_s^T \Lambda^1 \boldsymbol{e}_y + \boldsymbol{e}_t^T \Lambda^2 \boldsymbol{e}_y)\big)}{\sum_{y'} \exp\big(\mathbf{x}^T(\boldsymbol{e}_s^T \Lambda^1 \boldsymbol{e}_{y'} + \boldsymbol{e}_t^T \Lambda^2 \boldsymbol{e}_{y'})\big)} \quad (1)$$

where $\boldsymbol{a}^T \Lambda \boldsymbol{b}$ is a bilinear tensor product which results in a vector, $h \in \mathbb{R}^V$. We use this tensor-based notation mainly to facilitate the description of the generative training algorithm.

3.2 Generative Training

Consider the training dataset $\mathcal{D}_{train}$, containing instances of the binary feature vector $\mathbf{x}$. To train a generative model, we use minimization of the negative joint log-likelihood.

$$\mathcal{L}(\mathcal{D}_{train}) = -\sum_{i=1}^{|\mathcal{D}_{train}|} \log p(\mathbf{y}_{(i)}, \mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})$$

In order to minimize the negative log-likelihood, we would compute its gradient with respect to the model parameters. The exact gradient, for any parameter $\theta \in \{\Lambda^1, \Lambda^2\}$, can be written as,

$$\frac{\partial \log p(\mathbf{y}_{(i)}, \mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})}{\partial \theta} = -\mathbf{E}_{\text{data}}\left[\frac{\partial E(\mathbf{y}_{(i)}, \mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})}{\partial \theta}\right] + \mathbf{E}_{\text{model}}\left[\frac{\partial E(\mathbf{y}, \mathbf{s}, \mathbf{t}, \mathbf{x})}{\partial \theta}\right] \quad (2)$$

While the first expectation can be computed in closed form, the second expectation is intractable due to the partition function. However, we can approximate it by generating a sample from the underlying distribution estimated by an MCMC algorithm such as Gibbs sampling. But instead of running the Gibbs chain for the whole burn-in period, we can run the chain for only k steps. This approximation method is called k-contrastive divergence (CD) (Hinton, 2002), which can be interpreted as optimizing a difference of Kullback-Leibler divergences. The novelty of this approach is in setting the sampler's initial state for variables at a training sample $(\mathbf{y}_{(i)}, \mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})$; this way the energy surface is modified only around the data points. We used CD-1 in our work.

Given the conditional independence assertions and the binary features, it is straightforward to show[2],

$$p(\mathbf{x}|y, s, t) = \prod_k p(x_k = 1|y, s, t) = \prod_k \text{sigm}\big(\boldsymbol{e}_{\mathbf{x}_k}^T(\boldsymbol{e}_s^T \Lambda^1 \boldsymbol{e}_y + \boldsymbol{e}_t^T \Lambda^2 \boldsymbol{e}_y)\big)$$

where $\boldsymbol{e}_{\mathbf{x}_k}$ is the one-hot representation of feature $\mathbf{x}_k$. We note that if sentiment and target variables were treated as additional input variables and conditional independence was applied to them as well, this generative specification would become identical to *naive Bayes*, and no approximation would be necessary. It can also be shown that $p(y|\mathbf{x}, s, t)$ and other conditional distributions needed to perform Gibbs sampling, $p(s|\mathbf{x}, y)$ and $p(t|\mathbf{x}, y)$, all follow a softmax distribution (Salakhutdinov and Hinton, 2009). See Equation 1.

The gradient with respect to the l_{th} slice of the tensors Λ^1 and Λ^2 can be approximated by:

$$\frac{\partial E(\theta)}{\partial \Lambda^1_{[l]}} = -\sum_i (\mathbf{y}^l_{(i)} \boldsymbol{e}_{s_{(i)}} \mathbf{x}^T_{(i)} - \hat{\mathbf{y}}^l_{(i)} \hat{\mathbf{s}}_{(\mathbf{i})} \hat{\mathbf{x}}^T_{(i)}) \quad (3)$$

$$\frac{\partial E(\theta)}{\partial \Lambda^2_{[l]}} = -\sum_i (\mathbf{y}^l_{(i)} \boldsymbol{e}_{t_{(i)}} \mathbf{x}^T_{(i)} - \hat{\mathbf{y}}^l_{(i)} \hat{\mathbf{t}}_{(\mathbf{i})} \hat{\mathbf{x}}^T_{(i)}) \quad (4)$$

where $\hat{\mathbf{y}}_i$, $\hat{\mathbf{x}}_i$, $\hat{\mathbf{t}}_i$, and $\hat{\mathbf{s}}_i$ are samples from $p(y|\mathbf{x}, s, t)$, $p(\mathbf{x}|y, s, t)$, $p(t|\mathbf{x}, y)$, and $p(s|\mathbf{x}, y)$ after CD-1, respectively. However, we use a version of CD (Mnih and Hinton, 2007) in which, instead of sampling and obtaining a binary vector, we set $\hat{\mathbf{y}}_i$, $\hat{\mathbf{x}}_i$, $\hat{\mathbf{t}}_i$, and $\hat{\mathbf{s}}_i$ to the vector of probabilities given by the respective probability distributions.

3.3 Discriminative Training

To train a discriminative model, we minimize the *cross-entropy* error,

$$J(\theta) = -\sum_i^{|\mathcal{D}_{train}|} \sum_l \mathbb{1}(\mathbf{y}_{(i)} = l) \log p(l|\mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})$$

[2]For the sake of simplicity of presentation, we omit the bias units which are needed for generative training.

	Overall				Atheism	Climate	Feminism	Hillary	Abortion
Method	F_{favor}	$F_{against}$	$MicF_{avg}$	$MacF_{avg}$	F_{avg}	F_{avg}	F_{avg}	F_{avg}	F_{avg}
CNN	61.98	72.67	67.33	58.57	63.34	**52.69**	51.33	64.41	61.09
RNN	59.32	76.33	67.82	56.02	61.47	41.63	**62.09**	57.67	57.28
SVM	62.98	74.98	68.98	58.01	65.19	42.35	57.46	58.63	66.42
MaxEnt	60.78	73.41	67.10	56.37	60.82	41.43	55.73	59.87	63.99
NB	58.05	71.11	64.58	55.51	63.69	40.46	49.58	64.77	58.64
Disc-TS	62.96	76.12	69.55	58.85	61.90	41.73	56.76	63.91	**69.94**
Disc-STS	**64.43**	**77.62**	**71.03**	**61.40**	65.52	41.18	57.90	**74.48**	67.94
Gen-STS	61.43	77.02	69.23	60.41	**67.09**	50.04	53.77	71.25	59.92

Table 2: Results for Task A, reporting the official competition metric, overall $MicF_{avg}$, along with F_{avg} for each individual target, and the average of all individual F_{avg}, $MacF_{avg}$.

The gradients with respect to the l_{th} slice of the tensor Λ^1 and Λ^2 can be computed exactly,

$$\frac{\partial J(\theta)}{\partial \Lambda^1_{[l]}} = -\sum_i \boldsymbol{e}_{s_{(i)}} \mathbf{x}^T_{(i)} \big(\mathbb{1}(\mathbf{y}_{(i)} = l) - p(l|\mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})\big) \tag{5}$$

$$\frac{\partial J(\theta)}{\partial \Lambda^2_{[l]}} = -\sum_i \boldsymbol{e}_{t_{(i)}} \mathbf{x}^T_{(i)} \big(\mathbb{1}(\mathbf{y}_{(i)} = l) - p(l|\mathbf{s}_{(i)}, \mathbf{t}_{(i)}, \mathbf{x}_{(i)})\big) \tag{6}$$

In both discriminative and generative cases, the gradients can be regarded as update rules for the weights of the model, which are untied based on the sentiment (updates on Λ^1) or the target (updates on Λ^2) in the tweet.

4 Experiments

SemEval 2016 Task 6 (Mohammad et al., 2016) defined two stance classification tasks/datasets. The first one (Task 6.A) was a traditional supervised task, while the second one (Task 6.B) was a weakly supervised task wherein no tweet was stance-annotated. For both tasks, we used binary n-gram features: word n-grams (1–3 gram) and character n-grams (2–5 gram). We used $\ell 2$ regularization for our discriminative (Disc-STS) and generative (Gen-STS) models. For Task A, model hyper-parameters were estimated by cross-validation on the training set. For Task B, we used the dataset of the supervised task as the development set. Gen-STS was trained by stochastic gradient descent, and the learning rate was set to 0.0005. Disc-STS was trained in batch mode, and we used L-BFGS for optimization. Task B contains noisy stance labels; because of this, we performed early stopping in training to avoid overfitting to the wrong model. To this end, during parameter tuning on the development set, we used a larger range for the progress threshold in our grid search.

4.1 Supervised Task

SemEval-2016 Task 6.A provided stance-annotated tweets toward five targets: "Atheism", "Climate Change is a Real Concern", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion". The dataset contained 2,914 and 1,249 tweets for training and testing respectively.

4.1.1 Results

Table 2 shows the results for Task A. CNN (Wei et al., 2016) and RNN (Zarrella and Marsh, 2016) are convolutional neural network and recurrent neural network models that were the second best and the best system in the competition respectively. Both systems use pre-trained word embeddings before training for the task, which improves generalization and allows them to achieve good results on the task. The SVM classifier was the linear-kernel SVM used by task organizers, which was trained on the same features as ours (i.e., word n-grams (1–3 gram) and character n-grams (2–5 gram)). Two other reasonable baselines, which resemble our discriminative and generative models respectively, are *maximum entropy* (MaxEnt), and *naive Bayes* (NB) classifiers.

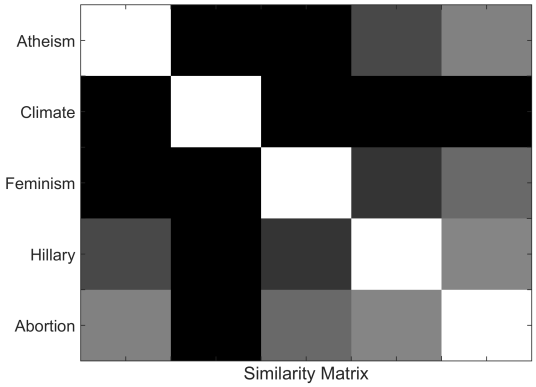

Figure 2: Similarity matrix of the weight vectors for task A targets. Lighter color denotes higher similarity.

The $MicF_{avg}$ metric is the mean of F_{favor} and $F_{against}$, which are the harmonic mean of Recall and Precision for each class. The metric $MicF_{avg}$ can be regarded as a micro-average of F-scores across targets. Alternatively, one could also determine the mean of the F_{avg} scores for each of the targets, the mean of which determines the ($MacF_{avg}$) metric.

Both Disc-STS and Gen-STS gain substantial gains over their natural baselines, MaxEnt and NB. Disc-STS improves the previous state-of-the-art results by 2.05% and 3.4% in Micro F1 and Macro F1 scores respectively. CNN and Gen-STS perform better on the "Climate" target, which is highly imbalanced (i.e., only 3.8% against). Unlike all the other baselines, which trained separate classifiers for each target, our approach can benefit from generalized features across multiple targets. Figure 2 displays the cosine similarity between the weight vectors for each of the targets. The weights used for this measure, were taken from the slice for that target, namely $\Lambda^2_{[target,:,:]}$.

Comparing MaxEnt and Disc-TS, the biggest improvement is found on the "Abortion" target. The performance on the targets, which are more similar to other targets in the corpus, is generally boosted significantly, compared with those that are not. The only difference between the two models is shared regularization across all the targets, which is causing the improvement.

Another way to investigate the inner workings of the model is to check if the model is able to discriminate sentiment features from target features. To do this, we represent the words based on the weights associated with them in the model. We concatenate the word-specific slices in the tensor parameter, namely $\Lambda^1_{[:,word,:]}$ and $\Lambda^2_{[:,word,:]}$, and compute the cosine similarity between pairs of word vectors. Table 3 shows the most similar words to 4 query words: two target-based and two sentiment-bearing words. It can be seen that among the top words similar to the sentiment-bearing words are some other sentiment-bearing words (positive or negative). The words similar to "climate" are clearly related to the target of "climate change". The words similar to "anti-choice" are about the target of "abortion", in addition to another related target, "feminism".

Given the significance of regularization and the dichotomy on the features, group lasso regularization (Yogatama and Smith, 2014), based on sentiment and target groups, can potentially improve our results.

We also report results on two subsets of the test set; (1) a subset where opinion is expressed toward the target; (2) a subset where opinion is expressed toward some other entity. Table 4 shows these results along with the overall $MicF_{avg}$, for the ease of reference.

4.2 Weakly Supervised Task

SemEval-2016 Task 6.B provided around 78,000 tweets associated with "Donald Trump". The tweets were gathered by polling Twitter for hashtags 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" and the use of other resources, e.g. lexicons, sentiment analyzers, etc., was permitted. This test set contained 707 tweets.

anti-choice	climate	excellent	crap
anti-abortion	global	together	asks
#feminism	co2	note	hack
effort	last	despite	hates
mentality	june	interesting	slut
benefits	warming	hate	shatter
unsafe	agriculture	retarded	#vaw
#reprorights	environmental	hatred	misogynistic
cunt	summer	warrior	adultery
types	mines	1st	either
banned	reducing	scum	societal

Table 3: Top similar words to 4 query words.

	Opinion toward		All
Method	Target	Other	
CNN	71.07	**46.66**	67.33
RNN	72.49	44.48	67.82
SVM	74.54	43.20	68.98
Disc-TS	74.60	44.95	69.55
Disc-STS	76.36	46.44	**71.03**
Gen-STS	**76.53**	43.39	69.23

Table 4: Results for Task A (the official competition metric F_{avg}) on different subsets of the test data.

4.2.1 Preprocessing

We only considered the tweets which contain no URL, are not retweets, are not shorter than 40 characters, and have at most three hashtags and three mentions. Following the protocol of the task, we start from labeling some hashtags. Among the most frequent hashtags in the training data, we manually labeled a handful of hashtags that are favorable to Trump, e.g., *#MakeAmericaGreatAgain*, and a handful of hashtags that are against him, e.g., *#TrumpYoureFired*. See Table 5 for a complete list of these hashtags. This weakly supervised approach gives us a dataset with *noisy* labels; for example, the tweet "*his #MakeAmericaGreatAgain #Tag is a bummer.*" is against Trump, incorrectly labeled favorable. Tweets that have at least one positive, or one negative hashtag/regex, and do not have both a positive and a negative hashtag/regex, are considered as our initial *favorable* and *against* instances. The final weakly labeled dataset consisted of a modest number of 1367 instances (544 against and 823 favorable).

We use a sentiment analyzer for tweets, VADER (Hutto and Gilbert, 2014), to classify the sentiments of the tweets. Here, we are dealing with only one target, but we still classify the tweets based on their topics. To do this, we use the standard topic modeling technique, LDA (Blei et al., 2003). This gives us an approximate fine-grained view of the topics of discussion in the data (e.g., immigration, Mexico, Obama, etc.). The number of topics (potential targets) was determined by the Elbow method (Thorndike, 1953), which was found to be 4. Finally, the topic distributions for the tweets were binarized (i.e., one for the dimension with the maximum value and zero for others).

4.2.2 Results

In Table 6 we compare our results with the best system in Task 6.B, which is the same CNN (Wei et al., 2016) system in Task 6.A, and state-of-the-art model, BiCond (Augenstein et al., 2016), which uses a bidirectional conditional LSTM encoding model. To handle the "neither" class we do the following: if the absolute value of the difference between the probability values of the two classes is less than a random number ($\epsilon|\epsilon \in (0, 0.1]$), then we classify it as "neither".

Figure 3 shows the impact of the amount of the training data on the performance of our models. Due to the limited nature of our data collection scheme, which tends to exploit only parts of the space of the

Favor.	#makeamericagreatagain, #illegalimmigration, #boycottmacys, #trumpisright, #trumpsright, #benghazi, #liberal-logic, #illegalimmigrant, #patriot, #standwithtrump, #leftists, #trumpfortriumph, #gotrump, #nobama
Against.	#gopclowncar, #racist, #hateisnotpresidential, #mexicanpride, #narcissist, #trumpsucks, #boycotttrump, #hishair, #proudlatina, #proudmexican, #trumpyourefired, #donaldtrumpsucks, #dumptrump, #partyofhate

Table 5: Stance-indicative hashtags used to collect favorable and opposing tweets.

Method	F_{favor}	$F_{against}$	F_{avg}
CNN	57.39	55.17	56.28
BiCond	**61.38**	54.68	**58.03**
Gen-STS	57.08	**56.38**	56.73
Disc-STS	39.59	55.43	47.51

Table 6: Evaluation on SemEval-2016 Task 6.B.

data, it is reasonable to expect that after a certain amount of data is seen, the performance of the system improves marginally as more training data is added. The discriminative model converges more quickly and performs poorly. Its performance improves marginally after seeing only 10% of the training data (i.e., 137 instances) and deteriorates soon. On the other hand, the generative model converges later with a much better F1 score. We also added 5% misclassification noise to the stance labels in the task A dataset but did not observe a similar pattern; instead, the discriminative classifier performed consistently better than the generative one and was less sensitive to the noise.

What we see in Figure 3 can be ascribed to the fragmentary view of the data created because of the hashtag-based process of bootstrapping a training set. In other words, the small number of tweets, which we harvest, covers only part of the test-data distribution. This is worsened by the lack of neutral tweets in the bootstrapped training set. Previous works have shown that variants of generative models alone, or their combination with discriminative models (Larochelle and Bengio, 2008; Nigam et al., 2000), are useful for classification especially when the amount of training data is limited (NG and Jordan, 2002). A detailed analysis of this phenomenon will be undertaken in the future.

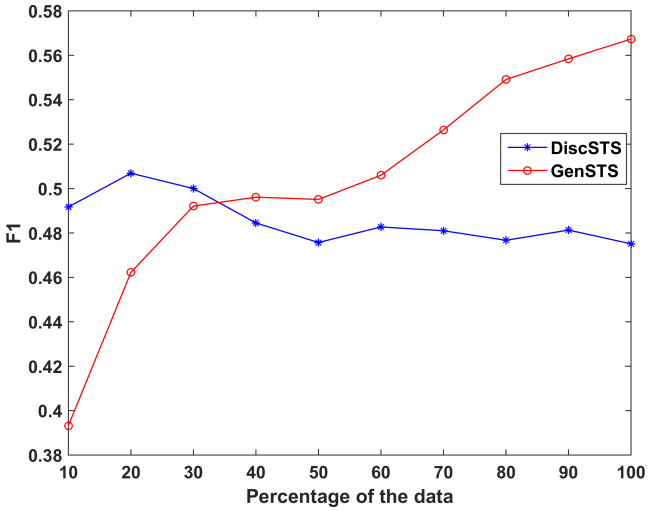

Figure 3: Comparison of GenSTS and DiscSTS on Task b. F1 is plotted against the amount of training data, i.e., percentage of the noisy-labeled data actually used for training. DiscSTS performs better initially and converges more quickly. GenSTS performs significantly better as more data is added.

5 Conclusion and Future Work

In this paper, we presented a log-linear approach for stance classification on tweets. The model employed sentiment and target variables in a novel way, wherein three-way interactions among input-sentiment-

stance variables and three-way interactions among input-target-stance variables were measured. Our findings show that the best way to use sentiments to improve stance classification is through these multi-way interactions. In addition, we demonstrated that by simply sharing regularization parameters among multiple targets, we are able to generalize features across multiple targets. While discriminative models are known to work better in classification tasks, generative models can also be useful when the data sample is small. Our results on a weakly labeled stance dataset proved that our generative model can in fact be much more effective than its discriminative counterpart.

For future work, our model can be easily incorporated in deep discriminative neural nets by replacing the standard softmax layer, effectively creating a multi-dimensional softmax layer. This has applications in tasks, wherein metadata exists; for example, a sentiment classification task for product reviews, in which metadata about the user and the products are also available. Moreover, the generative learning can be improved by replacing contrastive divergence with a more recent sampling method, SampleRank (Rohanimanesh et al., 2011), and using F1 score as the cost function.

6 Acknowledgement

This work was supported by NIH grant R01GM103309 and ARO grant W911NF-15-1-0265.

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 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of EMNLP*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Lipika Dey and SK Mirajul Haque. 2009. Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition*, 12(3):205–226.

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of EMNLP*.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. In *Proceedings of FLAIRS*, pages 174–179.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of NAACL-HLT*, pages 12–17.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of IJCNLP*, pages 1348–1356.

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.

Geoffrey Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

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.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of ACL*, pages 151–160.

Hugo Larochelle and Yoshua Bengio. 2008. Classification using discriminative Restricted Boltzmann Machines. In *Proceedings of ICML*, pages 536–543.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of ICML*, pages 641–648.

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 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.

Saif M Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of COLING*, pages 869–875.

Andrew Y NG and Michael I Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of NIPS*.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134.

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

Khashayar Rohanimanesh, Kedar Bellare, Aron Culotta, Andrew McCallum, and Michael L Wick. 2011. Samplerank: Training factor graphs with atomic gradients. In *Proceedings of ICML*, pages 777–784.

Ruslan R Salakhutdinov and Geoffrey E Hinton. 2009. Replicated softmax: an undirected topic model. In *Proceedings of NIPS*, pages 1607–1614.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceeding of the 2nd Workshop on Argumentation Mining*, pages 67–77.

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.

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.

Robert L Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18(4):267–276.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012a. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

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

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of SemEval*, pages 384–388.

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

Dani Yogatama and Noah A. Smith. 2014. Linguistic structured sparsity in text categorization. In *Proceedings of ACL*, pages 786–796.

Guido Zarrella and Amy Marsh. 2016. MITRE at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of SemEval*, pages 458–463.