Overcoming Language Variation in Sentiment Analysis with Social Attention

Yi Yang and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30308
{yiyang+jacobe}@gatech.edu

Abstract

Variation in language is ubiquitous, particularly in newer forms of writing such as social media. Fortunately, variation is not random; it is often linked to social properties of the author. In this paper, we show how to exploit social networks to make sentiment analysis more robust to social language variation. The key idea is *linguistic homophily*: the tendency of socially linked individuals to use language in similar ways. We formalize this idea in a novel attention-based neural network architecture, in which attention is divided among several basis models, depending on the author's position in the social network. This has the effect of smoothing the classification function across the social network, and makes it possible to induce personalized classifiers even for authors for whom there is no labeled data or demographic metadata. This model significantly improves the accuracies of sentiment analysis on Twitter and on review data.

1 Introduction

Words can mean different things to different people. Fortunately, these differences are rarely idiosyncratic, but are often linked to social factors, such as age (Rosenthal and McKeown, 2011), gender (Eckert and McConnell-Ginet, 2003), race (Green, 2002), geography (Trudgill, 1974), and more ineffable characteristics such as political and cultural attitudes (Fischer, 1958; Labov, 1963). In natural language processing (NLP), social media data has brought variation to the fore, spurring the development of new computational techniques for *characterizing* variation in the lexicon (Eisenstein et al., 2010), orthography (Eisenstein, 2015), and syntax (Blodgett et al., 2016). However, aside from the focused task of spelling normalization (Sproat et al., 2001; Aw et al., 2006), there have been few attempts to make NLP systems more robust to language variation across speakers or writers.

One exception is the work of Hovy (2015), who shows that the accuracies of sentiment analysis and topic classification can be improved by the inclusion of coarse-grained author demographics such as age and gender. However, such demographic information is not directly available in most datasets, and it is not yet clear whether predicted age and gender offer any improvements. On the other end of the spectrum are attempts to create *personalized* language technologies, as are often employed in information retrieval (Shen et al., 2005), recommender systems (Basilico and Hofmann, 2004), and language modeling (Federico, 1996). But personalization requires annotated data for each individual user—something that may be possible in interactive settings such as information retrieval, but is not typically feasible in natural language processing.

We propose a middle ground between group-level demographic characteristics and personalization, by exploiting social network structure. The sociological theory of *homophily* asserts that individuals are usually similar to their friends (McPherson et al., 2001). This property has been demonstrated for language (Bryden et al., 2013) as well as for the demographic properties targeted by Hovy (2015), which are more likely to be shared by friends than by random pairs of individuals (Thelwall, 2009). Social
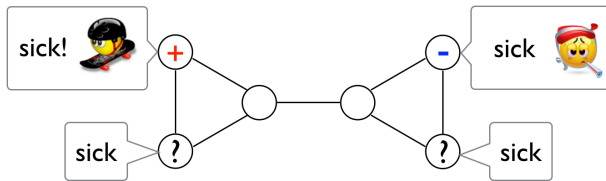
Figure 1: Words such as 'sick' can express opposite sentiment polarities depending on the author. We account for this variation by generalizing across the social network.

Dataset	# Positive	# Negative	# Neutral	# Tweet
Train 2013	3,230	1,265	4,109	8,604
Dev 2013	477	273	614	1,364
Test 2013	1,572	601	1,640	3,813
Test 2014	982	202	669	1,853
Test 2015	1,038	365	987	2,390

Table 1: Statistics of the SemEval Twitter sentiment datasets.

network information is available in a wide range of contexts, from social media (Huberman et al., 2008) to political speech (Thomas et al., 2006) to historical texts (Winterer, 2012). Thus, social network homophily has the potential to provide a more general way to account for linguistic variation in NLP.

Figure 1 gives a schematic of the motivation for our approach. The word 'sick' typically has a negative sentiment, e.g., 'I would like to believe he's sick rather than just mean and evil.'[1] However, in some communities the word can have a positive sentiment, e.g., the lyric 'this sick beat', recently trademarked by the musician Taylor Swift.[2] Given labeled examples of 'sick' in use by individuals in a social network, we assume that the word will have a similar sentiment meaning for their near neighbors—an assumption of *linguistic homophily* that is the basis for this research. Note that this differs from the assumption of *label homophily*, which entails that neighbors in the network will hold similar opinions, and will therefore produce similar document-level labels (Tan et al., 2011; Hu et al., 2013). Linguistic homophily is a more generalizable claim, which could in principle be applied to any language processing task where author network information is available.

To scale this basic intuition to datasets with tens of thousands of unique authors, we compress the social network into vector representations of each author node, using an embedding method for large

---
[1] Charles Rangel, describing Dick Cheney

[2] In the case of 'sick', speakers like Taylor Swift may employ either the positive and negative meanings, while speakers like Charles Rangel employ only the negative meaning. In other cases, communities may maintain completely distinct semantics for a word, such as the term 'pants' in American and British English. Thanks to Christopher Potts for suggesting this distinction and this example.

scale networks (Tang et al., 2015b). Applying the algorithm to Figure 1, the authors within each triad would likely be closer to each other than to authors in the opposite triad. We then incorporate these embeddings into an attention-based neural network model, called SOCIAL ATTENTION, which employs multiple basis models to focus on different regions of the social network.

We apply SOCIAL ATTENTION to Twitter sentiment classification, gathering social network metadata for Twitter users in the SemEval Twitter sentiment analysis tasks (Nakov et al., 2013). We further adopt the system to Ciao product reviews (Tang et al., 2012), training author embeddings using trust relationships between reviewers. SOCIAL ATTENTION offers a 2-3% improvement over related neural and ensemble architectures in which the social information is ablated. It also outperforms all prior published results on the SemEval Twitter test sets.

2 Data

In the SemEval Twitter sentiment analysis tasks, the goal is to classify the sentiment of each message as positive, negative, or neutral. Following Rosenthal et al. (2015), we train and tune our systems on the SemEval Twitter 2013 training and development datasets respectively, and evaluate on the 2013–2015 SemEval Twitter test sets. Statistics of these datasets are presented in Table 1. Our training and development datasets lack some of the original Twitter messages, which may have been deleted since the datasets were constructed. However, our test datasets contain all the tweets used in the SemEval evaluations, making our results comparable with prior work.

We construct three author social networks based on the follow, mention, and retweet relations between the 7,438 authors in the training dataset,

which we refer as FOLLOWER, MENTION and RETWEET.[3] Specifically, we use the Twitter API to crawl the friends of the SemEval users (individuals that they follow) and the most recent 3,200 tweets in their timelines.[4] The mention and retweet links are then extracted from the tweet text and metadata. We treat all social networks as undirected graphs, where two users are socially connected if there exists at least one social relation between them.

3 Linguistic Homophily

The hypothesis of *linguistic homophily* is that socially connected individuals tend to use language similarly, as compared to a randomly selected pair of individuals who are not socially connected. We now describe a pilot study that provides support for this hypothesis, focusing on the domain of sentiment analysis. The purpose of this study is to test whether errors in sentiment analysis are *assortative* on the social networks defined in the previous section: that is, if two individuals (i, j) are connected in the network, then a classifier error on i suggests that errors on j are more likely.

We test this idea using a simple lexicon-based classification approach, which we apply to the SemEval training data, focusing only on messages that are labeled as positive or negative (ignoring the neutral class), and excluding authors who contributed more than one message (a tiny minority). Using the social media sentiment lexicons defined by Tang et al. (2014),[5] we label a message as positive if it has at least as many positive words as negative words, and as negative otherwise.[6] The assortativity is the fraction of dyads for which the classifier makes two correct predictions or two incorrect predictions (Newman, 2003). This measures whether classification errors are clustered on the network.

We compare the observed assortativity against the assortativity in a network that has been randomly

rewired.[7] Each rewiring epoch involves a number of random rewiring operations equal to the total number of edges in the network. (The edges are randomly selected, so a given edge may not be rewired in each epoch.) By counting the number of edges that occur in both the original and rewired networks, we observe that this process converges to a steady state after three or four epochs. As shown in Figure 2, the original observed network displays more assortativity than the randomly rewired networks in nearly every case. Thus, the Twitter social networks display more linguistic homophily than we would expect due to chance alone.

The differences in assortativity across network types are small, indicating that none of the networks are clearly best. The retweet network was the most difficult to rewire, with the greatest proportion of shared edges between the original and rewired networks. This may explain why the assortativities of the randomly rewired networks were closest to the observed network in this case.

4 Model

In this section, we describe a neural network method that leverages social network information to improve text classification. Our approach is inspired by ensemble learning, where the system prediction is the weighted combination of the outputs of several basis models. We encourage each basis model to focus on a local region of the social network, so that classification on socially connected individuals employs similar model combinations.

Given a set of instances $\{\mathbf{x}_i\}$ and authors $\{a_i\}$, the goal of personalized probabilistic classification is to estimate a conditional label distribution $p(y \mid \mathbf{x}, a)$. For most authors, no labeled data is available, so it is impossible to estimate this distribution directly. We therefore make a smoothness assumption over a social network G: individuals who are socially proximate in G should have similar classifiers. This idea is put into practice by modeling the conditional label distribution as a mixture over the

---

[3]We could not gather the authorship information of 10% of the tweets in the training data, because the tweets or user accounts had been deleted by the time we crawled the social information.

[4]The Twitter API returns a maximum of 3,200 tweets.

[5]The lexicons include words that are assigned at least 0.99 confidence by the method of Tang et al. (2014): 1,474 positive and 1,956 negative words in total.

[6]Ties go to the positive class because it is more common.

[7]Specifically, we use the `double_edge_swap` operation of the `networkx` package (Hagberg et al., 2008). This operation preserves the degree of each node in the network.
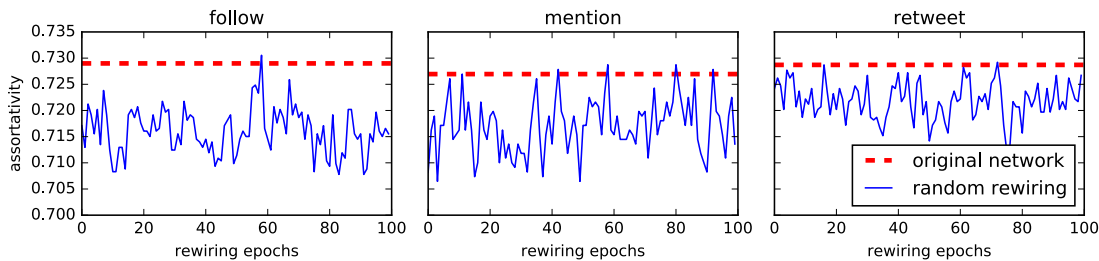
Figure 2: Assortativity of observed and randomized networks. Each rewiring epoch performs a number of rewiring operations equal to the total number of edges in the network. The randomly rewired networks almost always display lower assortativities than the original network, indicating that the accuracy of the lexicon-based sentiment analyzer is more assortative on the observed social network than one would expect by chance.

predictions of K basis classifiers,

$$p(y \mid \mathbf{x}, a) = \sum_{k=1}^{K} \Pr(Z_a = k \mid a, G) \times p(y \mid \mathbf{x}, Z_a = k). \tag{1}$$

The basis classifiers $p(y \mid \mathbf{x}, Z_a = k)$ can be arbitrary conditional distributions; we use convolutional neural networks, as described in § 4.2. The component weighting distribution $\Pr(Z_a = k \mid a, G)$ is conditioned on the social network G, and functions as an attentional mechanism, described in § 4.1. The basic intuition is that for a pair of authors a_i and a_j who are nearby in the social network G, the prediction rules should behave similarly if the attentional distributions are similar, $p(z \mid a_i, G) \approx p(z \mid a_j, G)$. If we have labeled data only for a_i, some of the personalization from that data will be shared by a_j. The overall classification approach can be viewed as a mixture of experts (Jacobs et al., 1991), leveraging the social network as side information to choose the distribution over experts for each author.

4.1 Social Attention Model

The goal of the social attention model is to assign similar basis weights to authors who are nearby in the social network G. We operationalize social proximity by embedding each node's social network position into a vector representation. Specifically, we employ the LINE method (Tang et al., 2015b), which estimates $D^{(v)}$ dimensional node embeddings $\mathbf{v}_a$ as parameters in a probabilistic model over edges in the social network. These embeddings are learned solely from the social network G, without leveraging

any textual information. The attentional weights are then computed from the embeddings using a softmax layer,

$$\Pr(Z_a = k \mid a, G) = \frac{\exp\left(\phi_k^\top \mathbf{v}_a + b_k\right)}{\sum_{k'}^{K} \exp\left(\phi_{k'}^\top \mathbf{v}_a + b_{k'}\right)}. \tag{2}$$

This embedding method uses only single-relational networks; in the evaluation, we will show results for Twitter networks built from networks of follow, mention, and retweet relations. In future work, we may consider combining all of these relation types into a unified multi-relational network. It is possible that embeddings in such a network could be estimated using techniques borrowed from multi-relational knowledge networks (Bordes et al., 2014; Wang et al., 2014).

4.2 Sentiment Classification with Convolutional Neural Networks

We next describe the basis models, $p(y \mid \mathbf{x}, Z = k)$. Because our target task is classification on microtext documents, we model this distribution using convolutional neural networks (CNNs; Lecun et al., 1989), which have been proven to perform well on sentence classification tasks (Kalchbrenner et al., 2014; Kim, 2014). CNNs apply layers of convolving filters to n-grams, thereby generating a vector of dense local features. CNNs improve upon traditional bag-of-words models because of their ability to capture word ordering information.

Let $\mathbf{x} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n]$ be the input sentence, where $\mathbf{h}_i$ is the $D^{(w)}$ dimensional word vector corresponding to the i-th word in the sentence. We use

298

one convolutional layer and one max pooling layer to generate the sentence representation of $\mathbf{x}$. The convolutional layer involves filters that are applied to bigrams to produce feature maps. Formally, given the bigram word vectors $\mathbf{h}_i, \mathbf{h}_{i+1}$, the features generated by m filters can be computed by

$$\mathbf{c}_i = \tanh(\mathbf{W}_L\mathbf{h}_i + \mathbf{W}_R\mathbf{h}_{i+1} + \mathbf{b}), \qquad (3)$$

where $\mathbf{c}_i$ is an m dimensional vector, $\mathbf{W}_L$ and $\mathbf{W}_R$ are $m \times D^{(w)}$ projection matrices, and $\mathbf{b}$ is the bias vector. The m dimensional vector representation of the sentence is given by the pooling operation

$$\mathbf{s} = \max_{i \in 1, \cdots, n-1} \mathbf{c}_i. \qquad (4)$$

To obtain the conditional label probability, we utilize a multiclass logistic regression model,

$$\Pr(Y = t \mid \mathbf{x}, Z = k) = \frac{\exp(\boldsymbol{\beta}_t^\top \mathbf{s}_k + \beta_t)}{\sum_{t'=1}^{T} \exp(\boldsymbol{\beta}_{t'}^\top \mathbf{s}_k + \beta_{t'})}, \qquad (5)$$

where $\boldsymbol{\beta}_t$ is an m dimensional weight vector, β_t is the corresponding bias term, and $\mathbf{s}_k$ is the m dimensional sentence representation produced by the k-th basis model.

4.3 Training

We fix the pretrained author and word embeddings during training our social attention model. Let Θ denote the parameters that need to be learned, which include $\{\mathbf{W}_L, \mathbf{W}_R, \mathbf{b}, \{\boldsymbol{\beta}_t, \beta_t\}_{t=1}^T\}$ for every basis CNN model, and the attentional weights $\{\phi_k, b_k\}_{k=1}^K$. We minimize the following logistic loss objective for each training instance:

$$\ell(\Theta) = -\sum_{t=1}^{T} \mathbf{1}[Y^* = t] \log \Pr(Y = t \mid \mathbf{x}, a), \quad (6)$$

where Y^* is the ground truth class for $\mathbf{x}$, and $\mathbf{1}[\cdot]$ represents an indicator function. We train the models for between 10 and 15 epochs using the Adam optimizer (Kingma and Ba, 2014), with early stopping on the development set.

4.4 Initialization

One potential problem is that after initialization, a small number of basis models may claim most of the mixture weights for all the users, while other basis

models are inactive. This can occur because some basis models may be initialized with parameters that are globally superior. As a result, the "dead" basis models will receive near-zero gradient updates, and therefore can never improve. The true model capacity can thereby be substantially lower than the K assigned experts.

Ideally, dead basis models will be avoided because each basis model should focus on a unique region of the social network. To ensure that this happens, we pretrain the basis models using an instance weighting approach from the domain adaptation literature (Jiang and Zhai, 2007). For each basis model k, each author a has an instance weight $\alpha_{a,k}$. These instance weights are based on the author's social network node embedding, so that socially proximate authors will have high weights for the same basis models. This is ensured by endowing each basis model with a random vector $\boldsymbol{\gamma}_k \sim N(\mathbf{0}, \sigma^2\mathbb{I})$, and setting the instance weights as,

$$\alpha_{a,k} = \text{sigmoid}(\boldsymbol{\gamma}_k^\top \mathbf{v}_a). \qquad (7)$$

The simple design results in similar instance weights for socially proximate authors. During pretraining, we train the k-th basis model by optimizing the following loss function for every instance:

$$\ell_k = -\alpha_{a,k} \sum_{t=1}^{T} \mathbf{1}[Y^* = t] \log \Pr(Y = t \mid \mathbf{x}, Z_a = k). \qquad (8)$$

The pretrained basis models are then assembled together and jointly trained using Equation 6.

5 Experiments

Our main evaluation focuses on the 2013–2015 SemEval Twitter sentiment analysis tasks. The datasets have been described in § 2. We train and tune our systems on the Train 2013 and Dev 2013 datasets respectively, and evaluate on the Test 2013–2015 sets. In addition, we evaluate on another dataset based on Ciao product reviews (Tang et al., 2012).

5.1 Social Network Expansion

We utilize Twitter's follower, mention, and retweet social networks to train user embeddings. By querying the Twitter API in April 2015, we were able

Network	# Author	# Relation
FOLLOWER+	18,281	1,287,260
MENTION+	25,007	1,403,369
RETWEET+	35,376	2,194,319

Table 2: Statistics of the author social networks used for training author embeddings.

to identify 15,221 authors for the tweets in the SemEval datasets described above. We induce social networks for these individuals by crawling their friend links and timelines, as described in § 2. Unfortunately, these networks are relatively sparse, with a large amount of isolated author nodes. To improve the quality of the author embeddings, we expand the set of author nodes by adding nodes that do the most to densify the author networks: for the follower network, we add additional individuals that are followed by at least a hundred authors in the original set; for the mention and retweet networks, we add all users that have been mentioned or retweeted by at least twenty authors in the original set. The statistics of the resulting networks are presented in Table 2.

5.2 Experimental Settings

We employ the pretrained word embeddings used by Astudillo et al. (2015), which are trained with a corpus of 52 million tweets, and have been shown to perform very well on this task. The embeddings are learned using the structured skip-gram model (Ling et al., 2015), and the embedding dimension is set at 600, following Astudillo et al. (2015). We report the same evaluation metric as the SemEval challenge: the average F1 score of positive and negative classes.[8]

Competitive systems We consider five competitive Twitter sentiment classification methods. *Convolutional neural network* (CNN) has been described in § 4.2, and is the basis model of SOCIAL ATTENTION. *Mixture of experts* employs the same CNN model as an expert, but the mixture densi-

---

[8]Regarding the neutral class: systems are penalized with false positives when neutral tweets are incorrectly classified as positive or negative, and with false negatives when positive or negative tweets are incorrectly classified as neutral. This follows the evaluation procedure of the SemEval challenge.

ties solely depend on the input values. We adopt the summation of the pretrained word embeddings as the sentence-level input to learn the gating function.[9] The model architecture of *random attention* is nearly identical to SOCIAL ATTENTION: the only distinction is that we replace the pretrained author embeddings with random embedding vectors, drawing uniformly from the interval $(-0.25, 0.25)$. *Concatenation* concatenates the author embedding with the sentence representation obtained from CNN, and then feeds the new representation to a softmax classifier. Finally, we include SOCIAL ATTENTION, the attention-based neural network method described in § 4.

We also compare against the three top-performing systems in the SemEval 2015 Twitter sentiment analysis challenge (Rosenthal et al., 2015): WEBIS (Hagen et al., 2015), UNITN (Severyn and Moschitti, 2015), and LSISLIF (Hamdan et al., 2015). UNITN achieves the best average F1 score on Test 2013–2015 sets among all the submitted systems. Finally, we republish results of NLSE (Astudillo et al., 2015), a non-linear subspace embedding model.

Parameter tuning We tune all the hyperparameters on the SemEval 2013 development set. We choose the number of bigram filters for the CNN models from {50, 100, 150}. The size of author embeddings is selected from {50, 100}. For *mixture of experts*, *random attention* and SOCIAL ATTENTION, we compare a range of numbers of basis models, {3, 5, 10, 15}. We found that a relatively small number of basis models are usually sufficient to achieve good performance. The number of pretraining epochs is selected from {1, 2, 3}. During joint training, we check the performance on the development set after each epoch to perform early stopping.

5.3 Results

Table 3 summarizes the main empirical findings, where we report results obtained from author embeddings trained on RETWEET+ network for SOCIAL ATTENTION. The results of different social networks for SOCIAL ATTENTION are shown in Table 4. The best hyperparameters are: 100 bigram

---

[9]The summation of the pretrained word embeddings works better than the average of the word embeddings.

System	Test 2013	Test 2014	Test 2015	Average
Our implementations				
CNN	69.31	72.73	63.24	68.43
Mixture of experts	68.97	72.07	64.28*	68.44
Random attention	69.48	71.56	64.37*	68.47
Concatenation	69.80	71.96	63.80	68.52
SOCIAL ATTENTION	71.91*	**75.07***	**66.75***	**71.24**
Reported results				
NLSE	72.09	73.64	65.21	70.31
WEBIS	68.49	70.86	64.84	68.06
UNITN	**72.79**	73.60	64.59	70.33
LSISLIF	71.34	71.54	64.27	69.05

Table 3: Average F1 score on the SemEval test sets. The best results are in **bold**. Results are marked with * if they are significantly better than CNN at $p < 0.05$.

	SemEval Test			
Network	2013	2014	2015	Average
FOLLOWER+	71.49	74.17	66.00	70.55
MENTION+	71.72	74.14	66.27	70.71
RETWEET+	**71.91**	**75.07**	**66.75**	**71.24**

Table 4: Comparison of different social networks with SOCIAL ATTENTION. The best results are in **bold**.

filters; 100-dimensional author embeddings; $K = 5$ basis models; 1 pre-training epoch. To establish the statistical significance of the results, we obtain 100 bootstrap samples for each test set, and compute the F1 score on each sample for each algorithm. A two-tail paired t-test is then applied to determine if the F1 scores of two algorithms are significantly different, $p < 0.05$.

Mixture of experts, random attention, and CNN all achieve similar average F1 scores on the SemEval Twitter 2013–2015 test sets. Note that random attention can benefit from some of the personalized information encoded in the random author embeddings, as Twitter messages posted by the same author share the same attentional weights. However, it barely improves the results, because the majority of authors contribute a single message in the SemEval datasets. With the incorporation of author social network information, concatenation slightly improves the classification performance. Finally, SOCIAL ATTENTION gives much better results than concatena-tion, as it is able to model the interactions between text representations and author representations. It significantly outperforms CNN on all the SemEval test sets, yielding 2.8% improvement on average F1 score. SOCIAL ATTENTION also performs substan-tially better than the top-performing SemEval sys-tems and NLSE, especially on the 2014 and 2015 test sets.

We now turn to a comparison of the social net-works. As shown in Table 4, the RETWEET+ net-work is the most effective, although the differences are small: SOCIAL ATTENTION outperforms prior work regardless of which network is selected. Twit-ter's "following" relation is a relatively low-cost form of social engagement, and it is less public than retweeting or mentioning another user. Thus it is unsurprising that the follower network is least useful for socially-informed personalization. The RETWEET+ network has denser social connections than MENTION+, which could lead to better author embeddings.

5.4 Analysis

We now investigate whether language variation in sentiment meaning has been captured by different basis models. We focus on the same sentiment words (Tang et al., 2014) that we used to test lin-guistic homophily in our analysis. We are inter-ested to discover sentiment words that are used with the opposite sentiment meanings by some authors. To measure the level of model-specificity for each

Basis model	More positive	More negative
1	banging loss fever broken <u>fucking</u>	**dear like** god **yeah wow**
2	chilling cold ill sick suck	satisfy trust wealth strong **lmao**
3	<u>ass</u> <u>damn</u> <u>piss</u> <u>bitch</u> <u>shit</u>	talent honestly voting win clever
4	insane bawling fever weird cry	**lmao** super **lol haha hahaha**
5	ruin silly bad boring dreadful	*lovatics* wish *beliebers arianators kendall*

Table 5: Top 5 more positive/negative words for the basis models in the SemEval training data. **Bolded** entries correspond to words that are often used ironically, by top authors related to basis model 1 and 4. <u>Underlined</u> entries are swear words, which are sometimes used positively by top users corresponding to basis model 3. *Italic* entries refer to celebrities and their fans, which usually appear in negative tweets by top authors for basis model 5.

Word	Sentiment	Example
sick	positive	Watch ESPN tonight to see me burning @user for a sick goal on the top ten. #realbackyardFIFA
bitch	positive	@user bitch u shoulda came with me Saturday sooooo much fun. Met Romeo santos lmao na i met his look a like
shit	positive	@user well shit! I hope your back for the morning show. I need you on my drive to Cupertino on Monday! Have fun!
dear	negative	Dear Spurs, You are out of COC, not in Champions League and come May wont be in top 4. Why do you even exist?
wow	negative	Wow. Tiger fires a 63 but not good enough. Nick Watney shoots a 59 if he birdies the 18th?!? #sick
lol	negative	Lol super awkward if its hella foggy at Rim tomorrow and the games suppose to be on tv lol Uhhhh.. Where's the ball? Lol

Table 6: Tweet examples that contain sentiment words conveying specific sentiment meanings that differ from their common senses in the SemEval training data. The sentiment labels are adopted from the SemEval annotations.

word w, we compute the difference between the model-specific probabilities $p(y \mid X = w, Z = k)$ and the average probabilities of all basis models $\frac{1}{K} \sum_{k=1}^{K} p(y \mid X = w, Z = k)$ for positive and negative classes. The five words in the negative and positive lexicons with the highest scores for each model are presented in Table 5.

As shown in Table 5, Twitter users corresponding to basis models 1 and 4 often use some words ironically in their tweets. Basis model 3 tends to assign positive sentiment polarity to swear words, and Twitter users related to basis model 5 seem to be less fond of fans of certain celebrities. Finally, basis model 2 identifies Twitter users that we have described in the introduction—they often adopt general negative words like 'ill', 'sick', and 'suck' positively. Examples containing some of these words are shown in Table 6.

5.5 Sentiment Analysis of Product Reviews

The labeled datasets for Twitter sentiment analysis are relatively small; to evaluate our method on a larger dataset, we utilize a product review dataset by Tang et al. (2012). The dataset consists of 257,682 reviews written by 10,569 users crawled from a popular product review sites, Ciao.[10] The rating information in discrete five-star range is available for the reviews, which is treated as the ground truth label information for the reviews. Moreover, the users of this site can mark explicit "trust" relationships with each other, creating a social network.

To select examples from this dataset, we first removed reviews that were marked by readers as "not useful." We treated reviews with more than three stars as positive, and less than three stars as negative; reviews with exactly three stars were removed.

[10]http://www.ciao.co.uk

Dataset	# Author	# Positive	# Negative	# Review
Train Ciao	8,545	63,047	6,953	70,000
Dev Ciao	4,087	9,052	948	10,000
Test Ciao	5,740	17,978	2,022	20,000
Total	9,267	90,077	9,923	100,000

Table 7: Statistics of the Ciao product review datasets.

System	Test Ciao
CNN	78.43
Mixture of experts	78.37
Random attention	79.43*
Concatenation	77.99
SOCIAL ATTENTION	**80.19****

Table 8: Average F1 score on the Ciao test set. The best results are in **bold**. Results are marked with * and ** if they are significantly better than CNN and random attention respectively, at $p < 0.05$.

We then sampled 100,000 reviews from this set, and split them randomly into training (70%), development (10%) and test sets (20%). The statistics of the resulting datasets are presented in Table 7. We utilize 145,828 trust relations between 18,999 Ciao users to train the author embeddings. We consider the 10,000 most frequent words in the datasets, and assign them pretrained word2vec embeddings.[11] As shown in Table 7, the datasets have highly skewed class distributions. Thus, we use the average F1 score of positive and negative classes as the evaluation metic.

The evaluation results are presented in Table 8. The best hyperparameters are generally the same as those for Twitter sentiment analysis, except that the optimal number of basis models is 10, and the optimal number of pretraining epochs is 2. Mixture of experts and concatenation obtain slightly worse F1 scores than the baseline CNN system, but random attention performs significantly better. In contrast to the SemEval datasets, individual users often contribute multiple reviews in the Ciao datasets (the average number of reviews from an author is 10.8; Table 7). As an author tends to express similar opinions toward related products, random attention

---

[11] https://code.google.com/archive/p/word2vec

is able to leverage the personalized information to improve sentiment analysis. Prior work has investigated the direction, obtaining positive results using speaker adaptation techniques (Al Boni et al., 2015). Finally, by exploiting the social network of trust relations, SOCIAL ATTENTION obtains further improvements, outperforming random attention by a small but significant margin.

6 Related Work

Domain adaptation and personalization Domain adaptation is a classic approach to handling the variation inherent in social media data (Eisenstein, 2013). Early approaches to supervised domain adaptation focused on adapting the classifier weights across domains, using enhanced feature spaces (Daumé III, 2007) or Bayesian priors (Chelba and Acero, 2006; Finkel and Manning, 2009). Recent work focuses on unsupervised domain adaptation, which typically works by transforming the input feature space so as to overcome domain differences (Blitzer et al., 2006). However, in many cases, the data has no natural partitioning into domains. In preliminary work, we constructed social network domains by running community detection algorithms on the author social network (Fortunato, 2010). However, these algorithms proved to be unstable on the sparse networks obtained from social media datasets, and offered minimal performance improvements. In this paper, we convert social network positions into node embeddings, and use an attentional component to smooth the classification rule across the embedding space.

Personalization has been an active research topic in areas such as speech recognition and information retrieval. Standard techniques for these tasks include linear transformation of model parameters (Legetter and Woodland, 1995) and collaborative filtering (Breese et al., 1998). These methods have recently been adapted to personalized sentiment analysis (Tang et al., 2015a; Al Boni et al., 2015). Supervised personalization typically requires labeled training examples for every individual user. In contrast, by leveraging the social network structure, we can obtain personalization even when labeled data is unavailable for many authors.

Sentiment analysis with social relations Previous work on incorporating social relations into sentiment classification has relied on the label consistency assumption, where the existence of social connections between users is taken as a clue that the sentiment polarities of the users' messages should be similar. Speriosu et al. (2011) construct a heterogeneous network with tweets, users, and n-grams as nodes. Each node is then associated with a sentiment label distribution, and these label distributions are smoothed by label propagation over the graph. Similar approaches are explored by Hu et al. (2013), who employ the graph Laplacian as a source of regularization, and by Tan et al. (2011) who take a factor graph approach. A related idea is to label the sentiment of individuals in a social network towards each other: West et al. (2014) exploit the sociological theory of structural balance to improve the accuracy of dyadic sentiment labels in this setting. All of these efforts are based on the intuition that individual predictions $p(y)$ should be smooth across the network. In contrast, our work is based on the intuition that social neighbors use language similarly, so they should have a similar *conditional* distribution $p(y \mid x)$. These intuitions are complementary: if both hold for a specific setting, then label consistency and linguistic consistency could in principle be combined to improve performance.

Social relations can also be applied to improve personalized sentiment analysis (Song et al., 2015; Wu and Huang, 2015). Song et al. (2015) present a latent factor model that alleviates the data sparsity problem by decomposing the messages into words that are represented by the weighted sentiment and topic units. Social relations are further incorporated into the model based on the intuition that linked individuals share similar interests with respect to the latent topics. Wu and Huang (2015) build a personalized sentiment classifier for each author; socially connected users are encouraged to have similar user-specific classifier components. As discussed above, the main challenge in personalized sentiment analysis is to obtain labeled data for each individual author. Both papers employ distant supervision, using emoticons to label additional instances. However, emoticons may be unavailable for some authors or even for entire genres, such as reviews. Furthermore, the pragmatic function of emoticons is com-

plex, and in many cases emoticons do not refer to sentiment (Walther and D'Addario, 2001). Our approach does not rely on distant supervision, and assumes only that the classification decision function should be smooth across the social network.

7 Conclusion

This paper presents a new method for learning to overcome language variation, leveraging the tendency of socially proximate individuals to use language similarly—the phenomenon of *linguistic homophily*. By learning basis models that focus on different local regions of the social network, our method is able to capture subtle shifts in meaning across the network. Inspired by ensemble learning, we have formulated this model by employing a social attention mechanism: the final prediction is the weighted combination of the outputs of the basis models, and each author has a unique weighting, depending on their position in the social network. Our model achieves significant improvements over standard convolutional networks, and ablation analyses show that social network information is the critical ingredient. In other work, language variation has been shown to pose problems for the entire NLP stack, from part-of-speech tagging to information extraction. A key question for future research is whether we can learn a socially-infused ensemble that is useful across multiple tasks.

8 Acknowledgments

References

Mohammad Al Boni, Keira Qi Zhou, Hongning Wang, and Matthew S Gerber. 2015. Model adaptation for

personalized opinion analysis. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Ramon F Astudillo, Silvio Amir, Wang Ling, Mário Silva, and Isabel Trancoso. 2015. Learning word representations from scarce and noisy data with embedding sub-spaces. In *Proceedings of the Association for Computational Linguistics (ACL)*.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Justin Basilico and Thomas Hofmann. 2004. Unifying collaborative and content-based filtering. In *Proceedings of the International Conference on Machine Learning (ICML)*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2014. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*.

John S Breese, David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*.

John Bryden, Sebastian Funk, and Vincent Jansen. 2013. Word usage mirrors community structure in the online social network twitter. *EPJ Data Science*, 2(1).

Ciprian Chelba and Alex Acero. 2006. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4).

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Penelope Eckert and Sally McConnell-Ginet. 2003. *Language and Gender*. Cambridge University Press.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19.

Marcello Federico. 1996. Bayesian estimation methods for n-gram language model adaptation. In *Proceedings of International Conference on Spoken Language (ICSLP)*.

Jenny R. Finkel and Christopher Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

John L Fischer. 1958. Social influences on the choice of a linguistic variant. *Word*, 14.

Santo Fortunato. 2010. Community detection in graphs. *Physics Reports*, 486(3).

Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

Aric A. Hagberg, Daniel A Schult, and P Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy)*.

Matthias Hagen, Martin Potthast, Michael Büchner, and Benno Stein. 2015. Webis: An ensemble for twitter sentiment detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Hussam Hamdan, Patrice Bellot, and Frederic Bechet. 2015. lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of Web Search and Data Mining (WSDM)*.

Bernardo Huberman, Daniel M. Romero, and Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1).

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1).

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

William Labov. 1963. The social motivation of a sound change. *Word*, 19(3).

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4).

Christopher J Leggetter and Philip C Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2).

Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation*.

Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review*, 45(2).

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and Post-Social media generations. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Aliaksei Severyn and Alessandro Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation*.

Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Implicit user modeling for personalized search. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*.

Kaisong Song, Shi Feng, Wei Gao, Daling Wang, Ge Yu, and Kam-Fai Wong. 2015. Personalized sentiment classification based on latent individuality of microblog users. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

R. Sproat, A.W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3).

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of Knowledge Discovery and Data Mining (KDD)*.

Jiliang Tang, Huiji Gao, and Huan Liu. 2012. mtrust: discerning multi-faceted trust in a connected world. In *Proceedings of Web Search and Data Mining (WSDM)*.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Duyu Tang, Bing Qin, and Ting Liu. 2015a. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015b. Line: Large-scale information network embedding. In *Proceedings of the Conference on World-Wide Web (WWW)*.

Mike Thelwall. 2009. Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2).

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2).

Joseph B. Walther and Kyle P. D'Addario. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review*, 19(3).

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Robert West, Hristo Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2.

Caroline Winterer. 2012. Where is America in the Republic of Letters? *Modern Intellectual History*, 9(03).

Fangzhao Wu and Yongfeng Huang. 2015. Personalized microblog sentiment classification via multi-task learning. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.