That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using *#petpeeve* Tweets *

William Yang Wang and **Diyi Yang**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
{yww,diyiy}@cs.cmu.edu

Abstract

We propose a novel data augmentation approach to enhance computational behavioral analysis using social media text. In particular, we collect a Twitter corpus of the descriptions of annoying behaviors using the *#petpeeve* hashtags. In the qualitative analysis, we study the language use in these tweets, with a special focus on the fine-grained categories and the geographic variation of the language. In quantitative analysis, we show that lexical and syntactic features are useful for automatic categorization of annoying behaviors, and frame-semantic features further boost the performance; that leveraging large lexical embeddings to create additional training instances significantly improves the lexical model; and incorporating frame-semantic embedding achieves the best overall performance.

1 Introduction

In the ever-expanding era of social media, many scientific disciplines, such as health and healthcare, biology, and learning sciences, have adopted computational approaches to exploit patterns and behaviors in large datasets (Wang et al., 2015; Chen and Lonardi, 2009; Baker and Yacef, 2009). In contrast, the primary methods for behavioral sciences still rely on lab experiments with limited amount of subjects, which are time consuming and financially expensive. In addition to this, it is also difficult to obtain a set of samples with geographical variations in traditional lab-based behavioral experiments.

I really hate being interrupted while I'm talking to someone #petpeeve #canyounot 🤦

Figure 1: An anonymized example of *#petpeeve* tweets.

While the social media data are abundantly available, computational approaches to behavioral sciences using Twitter are not well-studied. Even when statistical techniques are applied to these tasks, their concentration has been on simple statistical significance tests and descriptive statistics (De Charms, 2013; Zhang et al., 2013). Therefore, we believe that statistical natural language processing techniques are needed for insightful analysis and interpretation in behavioral studies.

In this paper, we use Twitter as a corpus for computational behavioral science. More specifically, we focus on a case study of analyzing annoying behaviors. To do this, we exploit a corpus of 9 million tweets (Cheng et al., 2010), and extract the tweets that describe these behaviors using the *#petpeeve* hashtags. *#petpeeve* is a popular Twitter hashtag, which describes behaviors that might be annoying to others. An example of *#petpeeve* tweets is shown in Figure 1. To facilitate the analysis, we manually annotate 3,375 tweets with 60 fine-grained categories, which will be described in Section 3. We use a sparse mixed-effects topic model to analyze the salient words in each category, as well as the geographic variations. We show that lexical, syntactic, and semantic features enhance the automatic categorization of annoying behaviors; and that the performance is further improved with a novel lexical and frame-semantic embedding based data augmentation ap-

*We understand that many people find long titles annoying, so we intentionally use a very long one to help people understand what "pet peeve" means.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2557–2563,
Lisbon, Portugal, 17-21 September 2015.

proach. Our main contributions are three-fold:

- We provide a Twitter corpus with fine-grained annotations for computational behavior studies;
- We qualitatively analyze the Twitter language concerning annoying behaviors, with a focus on the topics and geographical variations;
- We propose various linguistic features and a novel data augmentation approach for automatic categorization of annoying behaviors.

We outline related work in the next section. The dataset is described in Section 3. We introduce the approach for analyzing *#petpeeve* Tweets in Section 4. Experimental results are shown in Section 5. We discuss possible applications in Section 6, and conclude in Section 7.

2 Related Work

Psychologists, behavioral scientists, and computer scientists have studied a wide-range of methods for behavior extraction (Mast et al., 2015). For example, in lab experiments, arm and body postures (Marcos-Ramiro et al., 2013) are often used to extract self-touch and gestures, while eye gaze (Funes Mora and Odobez, 2012), head pose (Ba and Odobez, 2011), face location and motion (Nguyen et al., 2012), and full-body pose (Shotton et al., 2013) can also be used as cues to extract gazing, nodding, and arm-related behaviors. There are also significant amount of studies of extracting facial and speech features to understand smiling (Bartlett et al., 2008), eye contact (Marin-Jimenez et al., 2014), and verbal behaviors (Basu, 2002).

With the surge of interest in computational social science (Lazer et al., 2009), Twitter has become a popular resource to study data-driven methods in social science (Miller, 2011). For example, O'Connor et al. (2010a) align the Twitter messages with public opinion time series to study computational political science. Ritter et al. (2010) study Twitter dialogues using a clustering approach. Bollen et al. (2011) use a sentiment analysis approach to predict the American stock market via Twitter. Li et al. (2014b) have investigated the alignment of Twitter mood with weather for sentiment analysis. In recent years, language technology researchers have focused on developing genre-specific Twitter part-of-speech tagging (Gimpel et al., 2011), named entity recognition (Ritter et al., 2011), summarization (O'Connor et al., 2010b), sentiment analysis (Agarwal et al., 2011), event extraction (Ritter et al., 2012; Li et al., 2014a), paraphrasing (Xu et al., 2014), machine translation (Ling et al., 2013), and dependency parsing (Kong et al., 2014) methods. To the best of our knowledge, even though there have been studies on using Twitter hashtags to study language-related behaviors (González-Ibánez et al., 2011; Bamman and Smith, 2015), Twitter NLP approaches to non-linguistic behaviors are not well studied in general.

Label	%	Label	%
appearance	.14	services	.02
disrespect	.06	traffic	.02
language	.06	advertisement	.01
hygiene	.05	bragging	.01
relationship	.05	children	.01
dishonesty	.03	complaining	.01
hypocrisy	.03	indolence	.01
incompetence	.03	physical	.01
interruption	.03	punctuality	.01
monetary	.03	racial	.01
sexual	.03	religious	.01
arrogance	.02	selfishness	.01
celebrity	.02	silence	.01
ignorance	.02	smoking	.01
privacy	.02	talkative	.01
products	.02	weather	.01

Table 1: The categories and percentages of annoying behaviors in *#petpeeve* tweets in our dataset. Note that 17% of the *#petpeeve* tweets are identified as *other* unrelated behaviors (not shown).

3 The Dataset

We use the Twitter corpus with 9 million sampled messages collected in prior work (Cheng et al., 2010), which includes a total of 121K users. The dataset includes latitude and longitude information.

We extract 3,375 tweets[1] with *#petpeeve* hashtags. We follow past work to annotate the tweets (Ritter et al., 2012; Li et al., 2014a): we apply the *LDA clustering + human-identification* approach to label the categories of the described annoying behaviors in these tweets. The human annotation process includes two stages: first, the annotators identify the 50 categories from the clustering process, and use these topics as a candi-

[1] http://www.cs.cmu.edu/~yww/data/petpeeves.zip

date label set to annotate the data; in the second stage, the categories are refined (to 60 classes) from the first pass, and the data is re-annotated with the refined human-specified category labels. Due to the complexity of this fine-grained annotation task, the inter-annotator agreement rate between two annotators is moderate (0.445).

The annotated categories and label distribution[2] of the dataset are shown in Table 1. In our random samples, the states that post the most *#petpeeve* tweets are NY, MD, CA, NJ, FL, GA, VA, TX, NC, PA, and DC. In our predictive experiments, we randomly select 60% of tweets for training, and 40% for testing.

4 Our Approach

In this section, we describe our methods for the qualitative and quantitative analyses. In particular, we briefly review a supervised approach of using sparse mixed-effects topic model to visualize the topical words to analyze this behavior data. For the quantitative task of automatic categorization of tweets, we propose a novel approach to create additional training data, using continuous lexical and semantic representations.

4.1 Supervised Topic Modeling

To analyze the salient words for each category of annoying behaviors, we utilize SAGE (Eisenstein et al., 2011), a state-of-the-art mixed-effect topic model, which has been used in several NLP applications (Sim et al., 2012; Wang et al., 2012). SAGE is ideal for our text analytic purposes, because it is supervised, and it builds relatively clean topic models by considering the additive effects and the background distribution of words. Therefore, we can use SAGE to visualize the salient words for each category of annoying behaviors using the 3,375 *#petpeeve* tweets. Each tweet is treated as a document, and we use Markov Chain Monte Carlo for inference. To facilitate the geographical analysis, we use Google's reverse geocoding service to extract the state information from coordinates, and apply SAGE for visualization.

[2]The categories that are not shown in the table are *backstabbing, boring, copycat, drinking, drug, empty promise, impoliteness, inconsiderate, indirect, insecurity, interference, irresponsible, jealous, judge, loneliness, misunderstanding, negativity, noisy, parents, politics, repetition, showoff, snobbish, stability, swearing, time-wasting, ungratefulness, and others*.

4.2 Embedding-Based Data Augmentation for Automatic Categorization of Tweets

In addition to the visualization task, we also ask the question: can we use linguistic cues to predict tweets that describe different annoying behaviors? We formulate the problem as a multiclass classification task, and consider the following feature sets:

- **Lexical Features**: we extract unigrams as surface-level lexical features.
- **Part-of-Speech Features**: to model shallow syntactic cues, we extract lexicalized part-of-speech features using the Stanford part-of-speech tagger (Toutanova et al., 2003).
- **Dependency Triples**: to better understand the deeper syntactic dependencies of keywords in tweets, we have also extracted typed dependency triples (e.g., *nsubj(hate,I)*) using the MaltParser (Nivre et al., 2007).
- **Frame-Semantics Features**: SEMAFOR (Das et al., 2010) is a state-of-the-art frame-semantics parser that produces FrameNet-style semantic annotation. We use SEMAFOR to extract frame-level semantic features.

Embeddings for Data Augmentation Since the Twitter messages are often short and noisy, and the training data is relatively scarce for each class, we consider the feasibility of leveraging external resources, in particular, continuous word embeddings (Mikolov et al., 2013a) to enhance the multiclass text categorization model.

Two major challenges for leveraging word embeddings for tweet classification are: 1) because word embeddings are continuous, it is difficult to fuse them with other discrete syntactic and semantic features; 2) it is not straightforward how one should transform the word-level representation to the tweet-level representation. In our preliminary experiments, we have evaluated the continuous word representation method (Turian et al., 2010), as well as incorporating neighboring words in the embeddings as additional features, but both methods fail to outperform the lexical baseline that uses only bag-of-word unigrams.

To solve this problem, we propose the use of neighboring words in continuous representations to create new instances to augment the training

weather	**ungratefulness**	**traffic**	**timewasting**	**talkative**	**swearing**	**stability**	**snobbish**
rains	helped	cop	wastingmytime	Tweeters	curse	mood	smut
STORM	ungrateful	lane	colleagues	Xs	teary	sensitive	intellectual
Blizzarad	clearly	pulled	Wen	wht	qweet91	dudes	moneycars
snowed	r	speed	BruklynFinest	sheesh	swears	nigga	LoWQUI
SNOW	them	Slow	hold	TwitterJail	10	up	lifestyle
smoking	**silence**	**showoff**	**sexual**	**services**	**selfishness**	**repetition**	**religious**
JAYECANE	guilty	louis	box	fil	ONLY	dislike	sinners
reggie	R	rims	wonder	requests	Selfish	repeat	IAmKevinTerrell
smoking	response	seein	Preach	convos	selfish	myself	spiritual
smoke	conversation	makin	suck	TIP	stay	same	CHURCH
smokers	sending	bag	pussy	products	hit	over	FOLK

Table 2: The salient words for categories of annoying behaviors learned by the sparse additive generative model of text.

State	Top Topical Words
NY	stalkers niqqas der den part dats liek havin
MD	fuckouttahere missing ima dmv fan situation tongue
CA	pocket clown phones football fit acting lip
NJ	nite blame p hips pum summer elses seein
FL	daddy both chipped pum rims nappy foh children
GA	oo affioncrockett season cigarettes year tatoos
VA	lane language middle might check winter past duke
TX	drama lmaoooo gtfoh nappy two jk stare unfollow
NC	everyday ear chic during hello wayansjr tryn nicca
PA	10 huh killyaself lifestyle shades round texts fucc
DC	dmv uncle nosey stare cares bish 1st lips

Table 3: The geographical variation of the annoying behaviors.

dataset. More specifically, in the embedding vocabulary $\mathcal{W}$, we search for the k-nearest-neighbor (knn) word w for a query term using cosine similarity between query $\vec{Q}$ and target word vectors $\vec{W}$:

$$\arg\max_{w \in \mathcal{W}} cosine(\vec{Q}, \vec{W}) \quad (1)$$

For each word in a tweet, we query the external embeddings, and replace them with their knn words to create a new training instance. For example, consider the tweet *"Being late is terrible"* with the *punctuality* label, after searching for knn words for each token, we create a new training instance: *"Be behind are bad"* with the same label.

Frame-Semantic Embeddings Although lexical (Mikolov et al., 2013a) and dependency based embeddings (Levy and Goldberg, 2014) have been studied, semantic-based embedding is still less understood. We consider the continuous embedding of semantic frames (Baker et al., 1998). To do this, we semantically parsed 3.8 million tweets using SEMAFOR (Das et al., 2010), and built a continuous bag-of-frame model to represent each semantic frame using Word2Vec[3]. We then use the same data augmentation approach to create additional instances with these semantic frame embeddings.

[3]https://code.google.com/p/word2vec/

Features	Precision	Recall	F1
Lexical	.341	.342	.341
+POS	.345	.346	.346
+Dependency*	.349	.350	.350
+Semantic Frames*	**.365**	**.367**	**.366**

Table 4: Comparing linguistic features for categorizing annoying behaviors. The best results are highlighted in **bold**.* indicates that the result is significantly better than the lexical baseline ($p < .0001$).

5 Experiments

5.1 Qualitative Analysis

We show the results of the visualization of salient words for each category of tweets in Table 2. SAGE clearly does a good job identifying annoying specific behaviors in each category. For example, in the *traffic* category, we see that the keywords *"cop"* and *"pulled"* that associate with traffic stop are identified. Also, *"slow"* and *"speed"* are also recognized as annoying behaviors during traffic. In the *selfishness* category, the word *"ONLY"* and *"Selfish"* are corrected identified. In the *silence* category, we see that the word *"R"* is promising, because it indicates the behavior when someone reads a blackberry message without reply. We see that many slang expressions are associated with various labels.

In Table 3, we show the geographical variation of tweets. The word *"dmv"* (DC-Maryland-Virginia) is correctly associated with MD and DC, and when we search the database, these *#petpeeve* tweets mainly refer to the 2010 snowstorm in the Winter affecting these areas. The *"daddy"* is prominent in the state of Florida, while the word *"rims"* is also identified, showing the unique car culture of this southern state.

5.2 Quantitative Evaluation

Experimental Setup We use the logistic regression model from LibShortText (Yu et al., 2013)

Methods	Prec.	Rec.	F1	Imp.
Lexical Baseline (No Data Augmentation)	.341	.342	.341	—
+ UrbanDictionary Embeddings	.343	.344	.344	0.9%
+ Twitter Embeddings*	.357	.358	.358	4.7%
+ GoogleNews Embeddings*	**.364**	**.366**	**.365**	**6.1%**
All Features Baseline (No Data Augmentation)	.365	.367	.366	—
+ Lexical (GoogleNews) and Frame-Semantic Embeddings*	.376	.377	.376	2.7%
+ Lexical (Twitter) and Frame-Semantic Embeddings*	**.379**	.380	.379	3.6%
+ Lexical (UD) and Frame-Semantic Embeddings*	**.379**	**.381**	**.380**	**3.8%**

Table 5: The effectiveness of leveraging continuous embeddings to create additional training instances. Imp.: relative improvement to the baseline without data augmentation. The best results for each section are highlighted in **bold**.* indicates that the result is significantly better than the baseline without data augmentation ($p < .0001$).

as the classifier in our 60-way multi-class classification experiments. Grid search is used to select the best hyper-parameter using the training data only. A final classifier is then trained using the best hyper-parameters and test set results are reported. We set $k = 5$ for knn in our data augmentation experiments: the training data is expanded to 5 times of the original size. We use a paired two-tailed student's t test to assess the statistical significance.

Word2Vec is used to train various lexical and semantic embedding models. We consider three lexical embeddings and one frame-semantic embeddings for data augmentation: 1) GoogleNews Lexical Embeddings trained with 100 billion words (Mikolov et al., 2013b); 2) Twitter Lexical Embeddings trained with 51 million of words; 3) Urban Dictionary lexical embeddings trained with 53 million of words from slang definitions and examples; 4) Twitter Semantic Frame Embeddings trained with 27 million frames.

Varying Feature Sets We compare various features in Table 4. We see that adding shallow part-of-speech features does not have a strong effect on the performance, but adding the dependency triples significantly outperforms the lexical baseline. We see that the semantic frames are particular useful, showing a 7% relative improvement over the baseline.

The Effectiveness of Data Augmentation Table 5 shows the results of data augmentation. We see that using the Google News lexical embeddings to augment the training data brings a 6.1% relative F1 improvement over the lexical baseline. When considering the additional frame-semantic embeddings from Twitter, our system obtains the best F1 of 0.380, bringing a 3.8% improvement over the no data augmentation baseline with all linguistic features.

6 Discussion

We provide a case study of automatically categorizing annoying behaviors using *#petpeeve* Tweets. We hope that this study can further solicit relevant research on fine-grained analysis of annoying behaviors in different dimensions, and use computational approaches to improve social good. For example, by using coordinates and other APIs, one might analyze the annoying behaviors in the public working environments (e.g., office, meeting rooms, etc.). By understanding what annoys their employees, companies can renovate their working setups, refine their policies, and improve the satisfaction and productivity of their employees.

In addition to *#petpeeve* Tweets, there are many other interesting hashtags that align well with traditional topics in behavior sciences. For example, hashtags like *#occupywallstreet* can be used to study crowd behaviors in terms of a political unrest. The *#ALS* hashtag can be used to study public behaviors in reaction to philanthropic campaigns. Overall, Tweets from carefully selected hashtags can be inexpensive to obtain, and facilitate significant amount of behavioral studies.

7 Conclusion

In this paper, we have presented a case study of the annoying behaviors using Twitter as a corpus. Our fine-grained visualization approach shows insights of different categories of these behaviors, with the geographical effects. We also show that linguistic cues are useful to categorize these behaviors automatically, and that using lexical and semantic embeddings as a data augmentation method significantly improves the performance.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Sileye O Ba and Jean-Marc Odobez. 2011. Multiperson visual focus of attention from head pose and meeting contextual cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):101–116.

Ryan SJD Baker and Kalina Yacef. 2009. The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Marian Bartlett, Gwen Littlewort, Tingfan Wu, and Javier Movellan. 2008. Computer expression recognition toolbox. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–2. IEEE.

Sumit Basu. 2002. *Conversational scene analysis*. Ph.D. thesis, MaSSachuSettS InStitute of Technology.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Jake Y Chen and Stefano Lonardi. 2009. *Biological data mining*. CRC Press.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM.

D. Das, N. Schneider, D. Chen, and N.A. Smith. 2010. Probabilistic frame-semantic parsing. In *HLT-NAACL 2010*, page 948956, Los Angeles, California, USA, June.

Richard De Charms. 2013. *Personal causation: The internal affective determinants of behavior*. Routledge.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1041–1048.

Kenneth Alberto Funes Mora and J Odobez. 2012. Gaze estimation from multimodal kinect data. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 25–30. IEEE.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 581–586. Association for Computational Linguistics.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *EMNLP*.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721.

Omer Levy and Yoav Goldberg. 2014. Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014a. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jiwei Li, Xun Wang, and Eduard Hovy. 2014b. What a nasty day: Exploring mood-weather relationship from twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1309–1318. ACM.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen, and Daniel

Gatica-Perez. 2013. Body communicative cue extraction for conversational analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. 2014. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296.

Marianne Schmid Mast, Daniel Gatica-Perez, Denise Frauendorfer, Laurent Nguyen, and Tanzeem Choudhury. 2015. Social sensing for psychology automated interpersonal behavior assessment. *Current Directions in Psychological Science*, 24(2):154–160.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Greg Miller. 2011. Social scientists wade into the tweet stream. *Science*, 333(6051):1814–1815.

Laurent Nguyen, Jean-Marc Odobez, and Daniel Gatica-Perez. 2012. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 289–292. ACM.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010a. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010b. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proc of NAACL*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Yanchuan Sim, Noah A. Smith, and David A. Smith. 2012. Discovering factions in the computational linguistics community. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12 Special Workshop on Rediscovering 50 Years of Discoveries.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 740–749. Association for Computational Linguistics.

Shiliang Wang, Michael J Paul, and Mark Dredze. 2015. Social media as a sensor of air quality and public response in china. *Journal of medical Internet research*, 17(3).

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics (TACL)*, 2(1).

H Yu, C Ho, Y Juan, and C Lin. 2013. Libshorttext: A library for short-text classification and analysis. Technical report, Technical Report. http://www. csie. ntu. edu. tw/˜ cjlin/ papers/libshorttext. pdf.

Ni Zhang, Shelly Campo, Kathleen F Janz, Petya Eckler, Jingzhen Yang, Linda G Snetselaar, and Alessio Signorini. 2013. Electronic word of mouth on twitter about physical activity in the united states: exploratory infodemiology study. *Journal of medical Internet research*, 15(11).