Table 2: Statistics of Evaluation Data

	# of Tweets (Training)	# of Tweets (Test)
Positive instances	628	166
Negative instances	5,052	1,254
Total	5,680	1,420

models, we use the Twitter data set used in an ADR classification work [37]. The data set was created by first collecting tweets with generic and brand names of drugs (in a similar manner our *Model T-Drug* was created), and then a randomly selected sample of the data was annotated by two domain experts under the guidance of a pharmacology expert. The annotators had an inter-annotator agreement of 0.69 (Cohen's Kappa). The authors did not clarify if the annotations distinguished between ADR and ADE, but since ADE cases are a super-set of ADR cases,[2] we use this data set for evaluating our classification models. The same data set was also used in the PSB (Pacific Symposium on Biocomputing) 2016 Social Media Shared Task for ADR classification (Task 1).[1] The tweets in the data set have binary labels for ADR.

The original data set contained a total of 10,822 tweets. As Twitter's terms-of-service do not allow sharing of actual tweet text, the data set is only available via tweet IDs. At the time when we re-acquired the data using the IDs, only 7,100 (65.6%) tweets were still publicly available. We randomly divided the available tweets into training, validation and test data (60%-20%-20% split) for experiments and evaluation. We ensured that none of our unlabeled Twitter data that we used to train the semi-supervised models had any overlap with the evaluation data set (by checking for duplicate text content).

For semi-supervised classification, ConText's [11] default loss function optimizes Mean Squared Error with respect to all classification categories. Because we do a binary classification on a heavily imbalanced class distribution (11% positive and 89% negative), we needed to find an optimal threshold for the prediction probabilities when classifying tweets as ADE. For each of our models, we individually tune the threshold on the validation data so that F1-score for the ADE class is maximized. After the threshold parameter is tuned for each model, we combine the training and validation data to create a final training data set. Table 2 presents statistics of the final training (80%) and test (20%) data sets.

4. EVALUATION

To evaluate the performance of our semi-supervised CNN models, the trained models are applied to each tweet in the 20% held-out test data to predict a binary label (i.e., ADE or non-ADE). We use precision, recall and F1-score as the evaluation metrics and report the results for the ADE class. For comparison, we also present classification performance of a baseline method and a number of supervised classification models. The supervised models are trained on the same final training set (described in section 3.4.2) used in our semi-supervised models.

Drug-Health-Condition Baseline is a simple heuristic based prediction model that classifies a tweet as ADE if the tweet mentions both a drug name and a health condition. For the drug names, we use the dictionary we created from the Walgreen drugs list (described in Section 3.2.3) containing 380 commonly mentioned drug names in Twitter. For the health conditions, we use the dictionary we created (described in Section 3.2.4) containing 7,193 health conditions.

fastText is a fast algorithm recently released by researchers at Facebook [14] that efficiently learns text representations for classification. It achieves comparable to state-of-the-art performance on tasks such as sentiment analysis and tag prediction. The classifier does not use any domain or problem specific features, rather it uses bag-of-n-grams and transforms them into low dimensional vector space so that the features can be shared across classification categories despite their lexical differences. We train fastText on the final 80% training data (with the same vector dimension and epoch we use in our semi-supervised CNN models) and apply the trained model on the final 20% held-out test data for ADE classification.

Supervised CNN is a supervised convolutional neural network classifier trained only on labeled tweets. Since both the supervised and semi-supervised CNN models use the same labeled training instances, we compare their performances to determine if ADE classification results could be improved with additional unlabeled data. For training the supervised CNN, we use ConText with its default parameter settings.

ADR Classifier is a state-of-the-art binary classifier [37] that is designed to classify short texts into ADR or non-ADR categories. It is a supervised classifier that uses a wide range of features derived from n-grams, UMLS semantic types that represent medical concepts, phrases denoting change in sentiment, WordNet synsets, ADR lexicon, sentiment lexicon, topic model-based features, text length, presence of comparative/superlative, modals, etc. The authors reported achieving 53.8% F1-score when the model was trained and tested on the original Twitter data set (10,822 tweets). The authors further improved the classification performance up to 59.7% F1-score on the original Twitter data set by using 10,617 labeled posts from an online health community forum and 23,516 labeled medical case reports as additional training instances. Since it was not possible to re-create the exact training/test data due to unavailability of some tweets, we trained their classification model with our final training data set, and applied it to the tweets in our test data to predict ADE tweets.

4.1 Results

Table 3 presents our ADE classification results and comparisons. The simple *Drug-Health-Condition* baseline results have the highest recall (63.86%) among all classification results, but also the lowest precision (25.67%). This shows that a drug and a health condition co-occurring in a tweet does not necessarily indicate an ADE, since these tweets will also include *indication* events (the drug is used to treat the condition) among other possibilities. The low precision can also be attributed to the noise in the health conditions dictionary.

The next section in Table 3 shows the results for the supervised models. The fastText model, which is a general purpose text classification model, improves performance substantially (+12% F1-score) over the simple *Drug-Health-Condition* baseline, but at the cost of much lower recall. The supervised CNN improves performance further with a substantial recall gain, raising F1-score to 52.53%.

Table 3: ADE Tweet Classification Performance (P = Precision, R = Recall, F1 = F1-score, M = ∼1,000,000 tweets or sentences, K = ∼1,000 tweets or sentences), differences in performance from the last row, majority vote, are statistically significant over the non-shaded regions (p < 0.05)

Model	Unlabeled Data Size	P (%)	R (%)	F1 (%)
Baseline				
Drug-Health-Condition Baseline	NA	25.67	**63.86**	36.61
Supervised Models				
fastText [14]	NA	56.35	42.77	48.63
Supervised CNN	NA	55.33	50.00	52.53
ADRClassifier [37]	NA	55.63	53.62	54.60
Model Built with General Unlabeled Data				
Model 1 T-Random	43.5M	64.29	54.22	58.82
Selective Use of Unlabeled Data (individual data sets)				
Model 2 Sent-Health	484K	58.79	58.43	58.61
Model 3 T-Drug	157K	63.24	51.81	56.95
Model 4 T-Health-Condition	3.5M	64.67	58.43	61.39
Model 5 T-Health-Condition*	7M	65.73	56.63	60.84
Selective Use of Unlabeled Data (multiple data sets)				
Model 6 T-Drug-Condition-Sent-Health	4.2M	67.11	60.24	63.49
Model 7 T-Drug-Condition*-Sent-Health	7.7M	67.33	60.84	63.92
Ensemble Prediction				
Majority Vote	NA	**70.21**	59.64	**64.50**

The state-of-the-art ADR classifier [37] specifically designed for this task achieves the best results (54.60% F1-score) among the supervised models. Despite the difference with the original data set due to unavailability of some tweets, this is still comparable to the reported results in the original paper (53.8% F1-score when trained and tested with Twitter data).

Our results for *Model-1 T-Random*, which uses a massive collection of random Twitter data (43.5M), has substantial performance improvements across all three evaluation metrics when compared to the supervised CNN model, and also outperformed the ADR Classifier baseline. This demonstrates the merits of the semi-supervised CNN classification as it leverages unlabeled data for automatic feature learning, and provides a plausible alternative to supervised methods since human-annotated labeled data are costly to generate or rarely available. However, recall at 54.22%, although better than the ADR Classifier, can still be improved.

In the next section of Table 3, we present the classification results when unlabeled data are used selectively so that the semi-supervised models can learn the problem-specific region embeddings (i.e., for high-level problem concepts) effectively. These models (Model 2-5) use much less unlabeled data than *Model-1 T-Random* and their data set sizes range from only 484K to 7M. Their performances also vary as precision ranges from 58.79% to 65.73%, recall ranges from 51.81% to 58.43% and F1-score ranges from 56.95% to 61.39%. All of these models (Model 2-5) achieve better F1-score than the baseline method and supervised models, while the best performing model's F1-score (*Model-4 T-Health-Condition*) improves over *Model-1 T-Random*'s F1-score by +2.57%. The relative differences in the results for these models (Model 2-5) are hard to compare since their unlabeled data sizes differ from each other. We also do not expect these models to perform substantially better because, individually, their unsupervised learning was tailored to a specific type of high-level concept such as drug names or health conditions. Intuitively, to be able to classify ADE effectively, the model should learn region embeddings of all of these high-level concepts simultaneously.

Model 6 & 7 are semi-supervised CNN models that combine all of the unlabeled data used in Model 2-5, such that region embeddings for drug, health conditions and other possible high-level problem concepts can be learned together. Model-7 outperforms all the models that use individual data sets (Model 2-5), and improves F1-score by an additional +2.53%. This model uses about 7.7M instances, which is nearly 5.6 times fewer number of instances than the random tweets data, and yet, was able to increase F1-score by +5.1% (from *Model-1 T-Random*), reaching 63.92% F1-score. Previous best results [37] reported on this data set had 59.7% F1-score, where the classification model used annotated tweets, online health community forum posts and medical case reports —totalling nearly 40K labeled training instances, whereas in the semi-supervised framework, we could achieve improved classification performance on a comparable data set with only 5,680 labeled training instances and selective use of the unlabeled data.

Since each of our models learns the region embeddings from slightly different types of unlabeled data, in the final row of Table 3 we combine the predictions from different models (Model 2-7) using majority vote for an ensemble prediction method. We break ties using the best performing model's prediction (best F1-score on the validation data set). ADE tweet classification with this ensemble prediction achieves the highest precision (70.21%) and F1-score (64.50%) among all our experiments, outperforming the state-of-the-art supervised ADR classifier from previous work by +9.9% F1-score, while recall level remained similar or better than the individual models (Model 2-7). These final results are statistically significant (using paired bootstrap significance test [2]) for most of the precision and F1-score improvements (P < 0.05), with the exception of precision of Model 6 and F1-score of Model 6 & 7. The non-shaded regions in Table 3 indicate statistically significant differences from the *Majority Vote* prediction results.

Table 4: Examples of False Negatives, False Positives and Prediction Disagreement among Models

Example Tweet	Label	Prediction (Model 6)	Prediction (Model 7)	Majority Vote
False Negatives				
@USER And then I had *horrible sleep* once I took the *trazodone*. I just couldn't win, haha.	ADE	non-ADE	non-ADE	non-ADE
@USER: I run on Vyvanse and RedBull. So done with that life. *Vyvance cooked my brain like a stove top*	ADE	non-ADE	non-ADE	non-ADE
False Positives				
Doctor gave me *moxifloxacin to kill my sinus infection*. Bringing out the big guns; sick of being sick!	non-ADE	ADE	ADE	ADE
@USER depending on your pain, for me and *my back all over pain Cymbalta* has been a miracle *no side effects*	non-ADE	ADE	ADE	ADE
Prediction Disagreement among Models				
Medication side-affects have hit me hard today: I keep on *randomly jolting* or *rocking*, it is so bad I am finding it *hard to type*! *#Seroquel*.	ADE	non-ADE	ADE	ADE

4.2 Qualitative Analysis

Table 4 presents some example tweets for which our best models had prediction errors or there were prediction disagreements among the models.

False Negatives. In the first tweet in Table 4, the health condition 'horrible sleep' is another way of describing 'sleep disorder', but 'horrible sleep' was not in our health conditions dictionary. Both Models 6 & 7 could not recognize this tweet as ADE. The second example may refer to one of the mental side effects of the drug ('manic symptoms', 'bipolar illness', 'Seeing things or hearing voices that are not real', 'Believing things that are not true'), but described in a very informal way ('cooked my brain'). These examples demonstrate the challenge of detecting these medical concepts written in colloquial language.

False Positives. In the middle section of Table 4, we present two examples that our models falsely predicted as ADE. These examples describe an *indication* event where a drug is used to treat a condition. Even though the user specifically mentioned that the drug gives 'no side effect', the model may not have learned good region embeddings for such phrases.

Prediction Disagreement among Models. In the last section of Table 4, we show an example for which our models (model 6 and 7) had a disagreement in their predictions. Model 7 (T-Drug-Condition-Sent-Health) predicted it correctly as an ADE, but model 6 (T-Drug-Condition*-Sent-Health) misclassified it as non-ADE. 'keep on randomly jolting or rocking' and 'hard to type' are not in our health condition dictionary. It is possible that because of other health-related words in the expanded dictionary, Model 7 could recognize the tweet as ADE. The ensemble model that takes a majority vote could also predict the correct label.

5. CONCLUSION

We have presented our experiments with a semi-supervised CNN-based framework for classification of adverse drug events in tweets, and evaluated our models on the Twitter data set used in the PSB 2016 Social Media Shared Task. By leveraging different types of unlabeled data to learn phrase embeddings for the semi-supervised classification, our models outperformed a state-of-the-art ADE classification model by +9.9% F1-score. Our best model (ensemble prediction by majority vote) achieved 70.21% precision, 59.64% recall, and 64.50% F1-score, and to the best of our knowledge sets new state-of-the-art results for this data set. ADE classification in tweets can allow for early detection as a means to augment existing ADE surveillance systems, and our results suggest a feasible solution that does not require a large number of labeled instances. In future, we will explore automatic extraction of high-level ADE concept phrases with the help of learned region embeddings, to detect drug and side-effect associations. Removing noisy phrases from both drugs and health conditions dictionaries could be a possible improvement scope in future work. ADE extraction and colloquial language modeling are additional avenues worth exploring for improving ADE classification performance.

6. REFERENCES

[1] E. Benzschawel. Identifying potential adverse drug events in tweets using bootstrapped lexicons. Master's thesis, Brandeis University, 5 2016.

[2] T. Berg-Kirkpatrick, D. Burkett, and D. Klein. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[3] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

[4] E. G. Brown, L. Wood, and S. Wood. The medical dictionary for regulatory activities (meddra). *Drug Safety*, 20(2):109–117, 2012.

[5] H.-J. Dai, M. Touray, J. Jonnagaddala, and S. Syed-Abdul. Feature engineering for recognizing adverse drug reactions from twitter posts. *Information*, 7(2):27, 2016.

[6] C. N. dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, 2014.

[7] D. Egger, F. Uzdilli, M. Cieliebak, and L. Derczynski. Adverse drug reaction detection using an adapted sentiment classifier. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.

[8] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta. Digital drug safety surveillance: Monitoring

pharmaceutical products in twitter. *Drug Safety*, 37(5):343–350, 2014.

[9] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, pages 885 – 892, 2012.

[10] S. A. Hasan, Y. Ling, J. Liu, and O. Farri. Exploiting neural embeddings for social media data analysis. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, 2015.

[11] R. Johnson and T. Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In *Proceedings of the 29th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2015.

[12] R. Johnson and T. Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. *arXiv preprint arXiv:1602.02373*, 2016.

[13] J. Jonnagaddala, T. R. Jue, and H. Dai. Binary classification of twitter posts for adverse drug reactions. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Big Island, HI, USA*, 2016.

[14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[15] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[16] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, and J. A. Kors. Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1):1, 2014.

[17] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73 – 81, 2015.

[18] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[19] J. Lardon, R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, N. Texier, M. C. Jaulent, M. N. Beyens, A. Burgun, and C. Bousquet. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17(7):e171, 2015.

[20] R. Leaman, R. I. Dogan, and Z. Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[21] R. Leaman, R. Khare, and Z. Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, 2015.

[22] Y. Lecun and Y. Bengio. *Convolutional Networks for Images, Speech and Time Series*. The MIT Press, 1995.

[23] J. Y. Lee and F. Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[24] K. Lee, A. Agrawal, and A. Choudhary. Real-time disease surveillance using twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1474–1477, New York, NY, USA, 2013. ACM.

[25] K. Lee, A. Agrawal, and A. Choudhary. Mining social media streams to improve public health allergy surveillance. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 815–822, Aug 2015.

[26] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258, Dec 2011.

[27] N. Limsopatham and N. Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[28] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[30] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[31] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22:671–681, 2015.

[32] B. Ofoghi, S. Siddiqui, and K. Verspoor. Read-biomed-ss: Adverse drug reaction classification of microblogs using emotional and conceptual enrichment. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.

[33] V. Plachouras, J. L. Leidner, and A. G. Garrow. Quantifying self-reported adverse drug events on twitter: Signal and topic analysis. In *Proceedings of the 7th 2016 International Conference on Social Media & Society*, 2016.

[34] M. Rastegar-Mojarad, R. K. Elayavilli, Y. Yu, and H. Liu. Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. In *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.

[35] H. Sampathkumar, X. Chen, and B. Luo. Mining adverse drug reactions from online healthcare forums

using hidden markov model. *BMC Medical Informatics and Decision Making*, 14(1):1–18, 2014.

[36] A. Sarker, R. E. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212, 2015.

[37] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53:196 – 207, 2015.

[38] D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[39] Unified medical language system Umls metathesaurus. `https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`. [accessed September-2016].

[40] S. J. Yeleswarapu, A. Rao, T. Joseph, V. Saipradeep, and R. Srinivasan. A pipeline to extract drug-adverse event. *BMC Med. Inf. & Decision Making*, 14:13, 2014.

[41] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, 2015.

[42] Z. Zhang, J.-Y. Nie, and X. Zhang. An ensemble method for binary classification of adverse drug reactions from social media. *Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, 2016.