Assignment 1

Ravi Regulagedda

Machine Learning in CL

Preprocessing

Feature Selection using syntactic n-grams [1] This paper talks about the use of syntactic n-grams as features in a machine learning model. Having described n-grams and syntactic n-grams in detail, the authors go on to use them as features in the task of predicting author attribution. Having created features from data using sn-grams, the authors compared them with word-based n-grams, POS n-grams, and character-based n-grams for three classification models - Support Vector Machines, Naive Bayes and Decision Trees. Published in *Expert Systems with Applications, 2014* from Google Scholar

Neural Models for Text Normalization [2] This paper focuses the text normalization that has to be done before feeding the text data into a TTS (text-to-speech) model. The authors propose a neural network to treat this process of text normalization as a sequence-to-sequence problem. This approach has allowed the authors to also integrate tagging and segmentation into this process. They also use finite-state grammars to provide additional context to the neural net to improve performance. Published in *Computational Linguistics, vol 45, 2019* from Google Scholar

Word Embeddings and Machine Learning [3] This paper provides a new metric to identify the quality of the relationships between word embedding pairs (eg., king : man :: queen : woman). They show that averaging over multiple word pairs improves the quality as well as describing a new method to measure the quality of associations using cosine similarity. Published in *Proceedings of the 26th International Conference on Computational Linguistics, 2016* from Google Scholar

Applications

Phishing detection [4] This paper proposes a method which focuses on the semantic content of the message to detect whether it is a phishing attempt. The authors first propose extracting (verb-object) pairs by performing semantic analysis on the message text. This was fed into a Naive Bayes classifier which would generate a confidence score for those pairs to be phishing or not. Published in *12th IEEE International Conference on Semantic Computing, 2018* from Google Scholar

Humor Detection [5] The authors of this paper propose a method of automatically detecting 'funny and unusual' scientific papers. The first build a dataset of such papers (using their own criteria) and use it train a multi-layer perceptron. They then compared this approach with using BERT and SciBERT using features they defined on the same dataset. This approach proved superior in terms of accuracy. Published in *ACL — IJCNLP, 2021* via ACL Anthology.

Analysis of Schizophrenia in Reddit posts [6] The authors explore the linguistic markers of Schizophrenia through reddit posts. They collected and analyzed a large dataset of reddit posts/comments from people who *claimed* to have received a Schizophrenia diagnosis. They also trained a machine learning classifier on this data. Published in the *Proceedings of Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019* from the ACL Anthology.

Sarcasm Analysis [7] The authors propose and investigate different LSTM based NN models for detecting sarcasm from the conversational context taking into account the preceding and succeeding messages as part of the data for the classifier. It is tested on data from Twitter and discussion forums. Published in *Computational Linguistics, vol 44, 2018* via the ACL Anthology.

Irony Detection [8] The authors propose a method to detect irony in tweets using manually annotated phrases and a data-driven approach to augment the current irony (sentiment) detection models. Published in *Computational Linguistics, vol 44, 2018* via the ACL Anthology.

Miscellaneous

Transfer Learning in Natural Language Processing [9] This paper presents an approach for transfer learning in NLP and show how this can be integrated in downstream NLP tasks. Published in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019* via ACL Anthology.

Unsupervised Text Classification using Experts and Word Embeddings [10] This paper explores a method to categorize documents into classes in the vein of unsupervised learning. The method described focuses on finding the similarities in the text content of the documents. This method leverages the semantic and lexical similarities to perform this classification showing resultson par with supervised models. Published in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019* via ACL Anthology.

Commonsense Knowledge Mining from Pretrained Models [11] The authors propose a bi-directional pretrained model to perform commonsense data mining on a given corpus. They show that this model is not biased by prior data from the same problem and that this approach outperforms and generalizes better that supervised models for the same problem. Published in *EMNLP-IJCNLP, 2019* via ACL Anthology.

Event Extraction as Machine Reading [12] This paper models the event extraction problem as a combination of a supervised and an unsupervised step. Questions are generated form the text via an unsupervised method and BERT based question-answering process does the Event Extraction. The authors claim this enables them to strengthen the reasoning process of the Event Extraction. Published in *EMNLP, 2020*, via ACL Anthology.

A self-supervised method for Machine Translation [13] This paper presents an emergent Neural Machine Translation (NMT) method. This system simultaneously selects training data and learns the NMT representation. This is also self-supervised and performs comparatively with regard to other supervised techniques. Published in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* from ACL Anthology.

Modeling Asynchronous Conversations [14] The authors present a method for modeling asynchronous conversations (email etc). They propose an architecture based on LSTM-RNN which encodes the text to capture the conversational dependencies in the asynchronous context. They also adapt the model to learn from synchronous (real-time) contexts and use adversarial training to improve the performance of the model. Published in *Computational Linguistics, vol 44* via ACL Anthology.

Catastrophic Forgetting in Neural Machine Translation [15] The authors aim to resolve the problem of catastrophic forgetting in Neural Networks in the context of machine translation. They propose using Elastic Weight Consolidation which helps the model retain weights that might be lost otherwise when performing continual training on new data. The authors show that this approach beats state-of-the-art models which tend to fall prey to catastrophic forgetting. Published in *Proceedings of the 29th conference of ACL, 2019* via ACL Anthology.

Abusive Language Detection

Checking Hatespeech Detection Models [16] This paper specifies 29 model functionalities that can be used to test models of hatespeech to indetify model and data weaknesses and biases. Published in *ACL — IJCNLP, 2021* via ACL Anthology

Impact of Politically Biased Data on hatespeech classification [17] This paper aims to investigate the affects of dataset bias on hatespeech detection models. The authors build three different datasets with different political leanings (right, left and neutral) to show that the presence of a politial bias in the data impairs the performance of the datasets. Published in *Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020* from Google Scholar.

Automatic Harassment Classifier [18] The authors present a method for classification of harassing tweets by leveraging the targets of harassment to annotate the data. The posit that this approach would reduce misclassifications as the data any classifier learns from would itself be of a higher quality and more accurate. Published in *Proceedings of the Fourth Workshop on Online Abuse and Harms, 2020*, via ACL Anthology.

Two-Step approach for Abusive Language Classification [19] The authors explore a two-step method for classifying abusive language by first classifying into whether it is abusive or not and then classifying it further into different classes. They compare it with a direct multi-class classification model using a CNN achieving comparable performance in the two step method as in the state of the art single step classification. Published in *ArXiv*, from ArXiv.

Racial Bias in Datasets for Abusive Language detection [20] This paper compares results from five different Twitter datasets to show the difference in predictions of tweets written by African Americans in Standard American English vs African American Vernacular English (AAVE). All the models trained on all five datasets predict tweets in AAVE to be abusive showing the racial bias spread over these datasets. Any systems built on these datasets would have this bias written into it. Published in *ArXiv* from ArXiv.

Impact of Biased Datasets in Abusive Language Detection [21] This paper shows the influence of data bias in this task. The authors show the classification scores of many popular models were much lower once this bias was reduced by random sampling instead of focused sampling. Published in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019* via ACL Anthology.

Abusive Language Detection in Online Content [22] This paper proposes splitting the text into features built from n-grams, linguistic, syntactic and distributional semantics and training a classifier model on these. The authors are able to achieve good performance on unseen data by comparing their predicted class with labels assigned by 3 annotators on that same unseen data. Published in *WWW' 16, 2016*, via Google Scholar.

Class-based errors to detect out-of-vocabulary hate speech [23] The authors propose a novel method of training a 'predict the next character' model on the abusive data and then use the error of that model to feed into a neural network classifier. This way the error is used to inform the classifier. The authors show that it outperforms other text-categorizers in out-of-vocabulary instances. Published in *Proceedings of the First Workshop on Abusive Language Online, 2017*, via Google Scholar.

Annotator Influence on Hate Speech Detection [24] This paper examines the influence of the annotator on models trained on the datasets they annotate. The author finds that amateur annotators are more likely to label text as hate speech and and models trained on datasets annotated by experts outperform models trained on datasets by expert annotators. Published in *Proceedings of the first workshop on NLP and computational social science, 2016* via Google Scholar.

A lexicon for a feature based approach to detecting abusive words [25] The authors propose using features extracted from a manually defined base lexicon, which is then used to create a larger lexicon. They show that this lexicon can be used to build models which can detect abusive words in any context. Published in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018* via Google Scholar.

References

[1] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic n-grams as machine learning features for natural language processing," *Expert Systems with Applications*, vol. 41, no. 3, pp. 853–860, 2014. Methods and Applications of Artificial and Computational Intelligence.

[2] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural Models of Text Normalization for Speech Applications," *Computational Linguistics*, vol. 45, pp. 293–337, 06 2019.

[3] A. Drozd, A. Gladkova, and S. Matsuoka, "Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (Osaka, Japan), pp. 3519–3530, The COLING 2016 Organizing Committee, Dec. 2016.

[4] T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 300–301, 2018.

[5] C. Shani, N. Borenstein, and D. Shahaf, "How did this get funded?! Automatically identifying quirky scientific achievements," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 14–28, Association for Computational Linguistics, Aug. 2021.

[6] J. Zomick, S. I. Levitan, and M. Serper, "Linguistic analysis of schizophrenia in Reddit posts," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, (Minneapolis, Minnesota), pp. 74–83, Association for Computational Linguistics, June 2019.

[7] D. Ghosh, A. R. Fabbri, and S. Muresan, "Sarcasm analysis using conversation context," *Computational Linguistics*, vol. 44, pp. 755–792, Dec. 2018.

[8] C. Van Hee, E. Lefever, and V. Hoste, "We usually don't like going to the dentist: Using common sense to detect irony on Twitter," *Computational Linguistics*, vol. 44, pp. 793–832, Dec. 2018.

[9] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, (Minneapolis, Minnesota), pp. 15–18, Association for Computational Linguistics, June 2019.

[10] Z. Haj-Yahia, A. Sieg, and L. A. Deleris, "Towards unsupervised text classification leveraging experts and word embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 371–379, Association for Computational Linguistics, July 2019.

[11] J. Davison, J. Feldman, and A. Rush, "Commonsense knowledge mining from pretrained models," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 1173–1178, Association for Computational Linguistics, Nov. 2019.

[12] J. Liu, Y. Chen, K. Liu, W. Bi, and X. Liu, "Event extraction as machine reading comprehension," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 1641–1651, Association for Computational Linguistics, Nov. 2020.

[13] D. Ruiter, C. España-Bonet, and J. van Genabith, "Self-supervised neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 1828–1834, Association for Computational Linguistics, July 2019.

[14] S. Joty and T. Mohiuddin, "Modeling speech acts in asynchronous conversations: A neural-CRF approach," *Computational Linguistics*, vol. 44, pp. 859–894, Dec. 2018.

[15] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, "Overcoming catastrophic forgetting during domain adaptation of neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 2062–2068, Association for Computational Linguistics, June 2019.

[16] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 41–58, Association for Computational Linguistics, Aug. 2021.

[17] M. Wich, J. Bauer, and G. Groh, "Impact of politically biased data on hate speech classification," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, (Online), pp. 54–64, Association for Computational Linguistics, Nov. 2020.

[18] I. Arora, J. Guo, S. I. Levitan, S. McGregor, and J. Hirschberg, "A novel methodology for developing automatic harassment classifiers for Twitter," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, (Online), pp. 7–15, Association for Computational Linguistics, Nov. 2020.

[19] J. H. Park and P. Fung, "One-step and two-step classification for abusive language detection on twitter," 2017.

[20] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," 2019.

[21] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, "Detection of Abusive Language: the Problem of Biased Datasets," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 602–608, Association for Computational Linguistics, June 2019.

[22] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, (Republic and Canton of Geneva, CHE), p. 145–153, International World Wide Web Conferences Steering Committee, 2016.

[23] J. Serrà, I. Leontiadis, D. Spathis, G. Stringhini, J. Blackburn, and A. Vakali, "Class-based prediction errors to detect hate speech with out-of-vocabulary words," in *Proceedings of the First Workshop on Abusive Language Online*, (Vancouver, BC, Canada), pp. 36–40, Association for Computational Linguistics, Aug. 2017.

[24] Z. Waseem, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in *Proceedings of the first workshop on NLP and computational social science*, pp. 138–142, 2016.

[25] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words – a feature-based approach," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 1046–1056, Association for Computational Linguistics, June 2018.