Granular Emotion Detection in Social Media Using Multi-Discipline Ensembles

Robert H. Frye(✉) and David C. Wilson

University of North Carolina at Charlotte, Charlotte, NC 28223, USA
{rfrye13,davils}@uncc.edu
https://www.charlotte.edu/

Abstract. A variety of applications across industry and society have started to adopt emotion detection in short written text as a key enabling component. However, the task of detecting fine-grained emotions (e.g. love, hate, sadness, happiness, etc.) in short texts such as social media remains both challenging and complex. Particularly for high-stakes applications such as health and public safety, there is a need for improved performance. To address the need for more accurate emotion detection in social media (EMDISM), we investigated the performance of ensemble classification approaches, which combine baseline models from machine learning, deep learning, and transformer learning. We evaluated a variety of ensemble approaches in comparison to the best individual component model using an EMDISM Twitter dataset with more than 1.2M samples. Results showed that the most accurate ensemble approaches performed significantly better than the best individual model.

Keywords: Emotion detection · Sentiment analysis · Social media · Ensemble · Transformer learning · Machine learning · Deep learning

1 Introduction

Understanding a person's emotional context by way of sentiment analysis or finer-grained emotion detection from written text can play a significant role in intelligent systems and modern applications, such as in commercial, political, or security areas [50]. Sentiment analysis (SA) is an application of Natural Language Processing (NLP) focused on determining the polarity of emotions in a textual or spoken sample (i.e., positive, negative, neutral). On a finer-grained level, emotion detection refines the task of sentiment analysis into classifying a sample as representative of specific emotions (e.g., happy, sad, angry, etc.). Illustrative commercial applications include identifying angry customers based on email content [23] as well as proper routing and escalation of messages to appropriate customer representatives [28].

Correctly identifying specific emotions in written text is challenging, even with richer data where texts are longer and well-written stylistically. However, texts in modern communication are more often aligned in structure with social

M. Ceci et al. (Eds.): ISMIS 2022, LNAI 13515, pp. 3–12, 2022.
https://doi.org/10.1007/978-3-031-16564-1_1

media interactions (shorter, less formally written), which present even greater challenges. *Emotion detection in social media* (EMDISM) must consider the less formal nature of the communication medium, with little regulation of writing styles and generally smaller sample sizes for analysis [21]. EMDISM is important for a variety of application contexts. For example, marketers and airlines apply sentiment analysis or EMDISM to assess emotional responses to advertising and understand overall customer satisfaction with travel experiences based on social media posts [24,43,47]. Beyond commercial applications, mental health providers monitor social media to identify indicators of depression [14], and security researchers are working to identify emerging threats from extremists [3] and other violent actions [34] from social media posts. Developing improved EMDISM approaches is broadly important for industry and society, and improving accuracy is a key open research question.

Our research is focused on the potential for improving accuracy in EMDISM applications by investigating ensemble approaches. In this paper, we present an in-depth evaluation of ensemble EMDISM approaches combining 15 common classifiers from 3 classification disciplines in 21 unique combinations across 4 categories of ensembles. We discuss key design decisions and experimental results indicating which ensembles were more effective than singleton classifiers and present significance testing demonstrating ensembles are often more accurate than singleton classifiers.

2 Related Work

In previous related research, we characterize three primary types of approach for sentiment analysis and emotion detection: machine learning (ML), deep learning (DL), and transformer learning (TL). Our research focuses on creating ensembles comprised of ML, DL, and TL classifiers, which have been previously applied to the tasks of text-based sentiment analysis or EMDISM. We present background research on individual component DL, ML, and TL classifiers, as well as ensemble approaches for leveraging combinations of component classifier outcomes.

2.1 Classifiers

Traditional **machine learning** (ML) classifiers generally apply logic or statistical analysis for text classification, and were among the earliest text classification algorithms. *Decision trees* have been applied to numerous classification problems, including EMDISM [36], and are a type of supervised learning algorithm, which builds classification structures based on partitioning data into subsets of samples with similar characteristics. Decision trees are one of the easiest classification methods for humans to understand, as they can be presented as graphs resembling trees, where each branch is a decision point and each leaf is a classification node. Ranganathan [36] applied decision trees to Twitter EMDISM of five emotions with reported accuracies between 88% to 96%. *Support vector machine* (SVM) [41] classifiers attempt to define a theoretical hyperplane used to segregate large vectors of sparsely populated data into discrete clusters with maximized distances between clusters, and given the sparse vector representations generated through tokenization of text. SVM has been widely applied to SA and EMDISM [8,32]. *Support vector classification* (SVC) [16] is used for processing high dimensional sparse vectors by "...reducing the number of objects in the training set that are used for defining the classifier." *LinearSVC* [46] is a variant of SVC designed to better scale to larger datasets. *Logistic regression* [18] uses independent variables to predict between binary classes, and has been applied in a one versus rest approach for SA and EMDISM [35].

Deep learning (DL) classifiers utilize layered neural networks and backwards propagation of error correction to create class predictions from tokenized embedding layers. DL classifiers for text classification generally consist of an embedding layer of tokenized text data, one or more hidden layers of decision neurons, and an output layer for predicting sample classes [52]. Complex neural networks have been developed, including *convolutional neural networks* (CNN) [30], which establish progressively smaller filters on samples to retain data about the context of one token to other tokens around it, and *recurrent neural networks* (RNN) which use an internal memory of previous steps to preserve contextual information about the relationships between tokens. *Bidirectional RNNs* (B-RNN) [38] and *long short-term memory* (LSTM) [26] neural networks were adapted versions of RNNs designed to address the vanishing or exploding gradient problem. B-RNNs use stacked RNNs to capture the context before and after a token, by training one RNN with tokens in the original order and the other RNN with tokens in reverse order. LSTM uses a combined forget gate, input gate, hidden memory layer, and output gate at each time step in the training process, and several variations of LSTM have been created including *gated recurrent units* (GRU) [11], *bidirectional GRU* (BiGRU) [10], *bidirectional LSTM* (BiLSTM) [39], and *convolutional LSTM* (C-LSTM) [22]. GRU combines LSTM's input and forget gates and merges the hidden memory layer and cell states, and BiGRU and BiLSTM add a bidirectional layer to GRU and LSTM respectively. C-LSTM adds memory of the class label to each gate in the LSTM layer.

Transformer learning (TL) classifiers, first proposed by Vaswani et al. [42], are a specific type of neural network which replace the convolutions and recurrence of DL classifiers with a paired encoder and decoder and a self-attention mechanism, which combine to effectively capture the context of each token in relation to other tokens in each sample. As TL classifiers avoid the need for recurrence or convolution, they generally require fewer epochs to fine-tune their base models and are more accurate than DL classifiers. *BERT* (Bidirectional Encoder Representations from Transformers) was developed by Devlin et al. [15] and used a masked language model approach to train their base model. BERT achieved an SST-2 accuracy score for the GLUE benchmarks [44] of 91.6% for binary SA. *RoBERTa* [31] attempted to improve upon BERT by training by training with larger batch sizes, more training epochs, and a larger vocabulary. RoBERTa achieved an SST-2 accuracy of 92.9%. *XLNet* [49] avoids the introduction of noise caused by inserting masking and separator tokens during BERT pre-training, and also considers permutations of factorization orders to

Table 1. Ensembles applied to text-based sentiment analysis or emotion detection.

Approach	Ensemble components	Type	Metric	Score
Kang et al. (2018) [27]	Hidden Markov Models	SA	Acc.	86.10%
Xia et al. (2011) [48]	NB, SVM, Maximum Entropy	SA	Acc.	80%–88%
Da Silva et al. (2014) [13]	NB, SVM, Random Forest, Logistic Regression, Lexicon	SA	Acc.	70%–79%
Araque et al. (2017) [2]	NB, ME, SVM, RNN, Lexicon	SA	Acc.	85%–94%
Perikos et al. (2016) [33]	NB, ME, knowledge-based	SA	Acc.	89%
Baziotis et al. (2018) [4]	Bi-directional LSTM ensemble	SA-Irony	Acc.	78.50%
Cao and Zukerman (2012) [9]	Lexicon-based, NB, ensemble SVM	ED-5 star	Acc.	70%–75%
Duppada et al. (2018) [17]	XG Boost and Random Forest	ED-4 class	Acc.	83.60%
Bickerstaffe et al. (2010) [5]	SVM, Decision Trees	ED-4 star	Acc.	49%–76%
Al-Omari et al. (2019) [1]	Fully connected NN, LSTM	ED-4 class	F1	0.67
Yue et al. (2018) [51]	CNN, RCNN, LSTM	ED-5 class	F1	0.468

capture the bidirectional context of tokens and maximize the probability that a token sequence would be present in each permutation. XLNet was 94.4% accurate in the SST-2 task. Lample and Conneau [29] developed the cross lingual model, *XLM*, to extend the concepts of BERT to additional languages, using 7500 training samples from 15 languages. *XLM-RoBERTa* (XLM-R) integrated concepts from XLM and BERT by applying MLM training with a larger vocabulary consisting of 250K tokens from 100 different languages compared to the 30K vocabulary used for BERT. XLM-R reported 95.0% accuracy in the SST-2 task. Clark et al. presented *ELECTRA* (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [12], which was designed to offset an imbalance caused by introducing masked tokens during pre-training BERT base models but not during fine-tuning. ELECTRA delivered SST-2 accuracy between 89.1% and 96.7% depending on training duration and which dataset was used for fine-tuning.

2.2 Ensembles

Ensemble classifiers are designed to offset the weaknesses of one or more classifiers with the strengths of other classifiers. Hansen and Salamon [25] suggested ensembles can be more accurate than singleton classifiers and that the correct first step for creating ensembles was to assess individual classifiers for accuracy to determine their suitability for inclusion in an ensemble. Boosting [37] is a process whereby iterative training and adjusting of weights is used to turn weak classifiers into strong classifiers, and AdaBoost [20] uses a weighted voting ensemble which is still in popular use. Bootstrap aggregating (bagging) [6] concepts included simple voting among base learners trained on different replicas of data, and this ensemble voting approach is still in use for SA and emotion detection today [4,32,48]. Burke [7] described numerous architectures for creating hybrids (ensembles) for recommender systems, including weighted voting, cascading, and switching approaches, among others. We adopt Burke's characterizations in discussing our ensemble approaches. Several research teams have created and applied ensembles combining various classifiers for sentiment analysis or emotion detection. Table 1 provides a list of ensemble researchers, the ensemble components they assessed, and the metrics reported for each approach [1,2,4,5,9,13,17,27,33,48,51]. Previous ensemble research has generally focused on binary sentiment analysis or classifying a more limited sampling of emotions with one of a few classifiers, whereas we have developed and assessed ensembles to classify a larger number of emotions (7) developed from a broader, cross-disciplinary selection of ML, DL, and TL classifiers.

3 Ensemble Approach and Evaluation

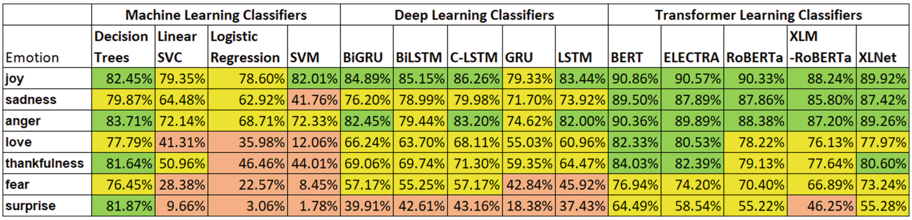
The specific challenge our research addresses focuses on potential performance improvements in finer-grained emotion detection in social media text. To address this challenge, we investigated the potential of ensemble approaches to improve performance in EMDISM. We conducted an in-depth evaluation of ensemble EMDISM approaches combining 15 common classifiers from 3 classification disciplines in 21 unique combinations across 4 categories of ensembles.

3.1 Experimental Setup

Our experiments were completed on a Micro-star International Z390 Gaming Infinite X Plus 9 desktop computer, with 48 GB of RAM, an Intel(R) Core(TM) i7-9700K CPU, and one NVIDIA GeForce RTX 2080 GPU. Our experimental platform was created in Python—using the Scikit-learn library for ML models, partitioning training/testing data, and analyzing results; Keras Tensorflow for DL model creation; the HuggingFace's Transformers and Simple Transformer libraries for TL model fine-tuning; Pandas and Numpy for dataframe and array processing; and NLTK for preprocessing text. We selected the EMDISM dataset developed by Wang et al. [45], hereafter referenced as the *HT* dataset. The HT dataset originally consisted of 2.5M Twitter tweets labeled with seven emotions—**joy**, **sadness**, **anger**, **love**, **thankfulness**, **fear**, and **surprise**—which are closely aligned with Ekman's six basic emotions [19]. At the time of our experimentation, the text detail of only 1.2M HT tweets remained available for hydration from Twitter with 349,419 samples of joy, 299,412 of sadness, 261,806 of anger, 153,017 of love, 72,505 of thankfulness, 65,010 of fear, and 11,978 of surprise. We followed common pre-processing steps [39,40] to de-noise the dataset. Specifically, we removed URLs, usernames, hashtags, and numbers, cast all text to lowercase, un-escaped html escape strings, replaced duplicate punctuation with singles (e.g. !!! became !), stripped extra whitespace, and lemmatized verbs. For experimentation, we performed 10-fold cross-validation testing and compared validation loss and accuracy curves to avoid overfitting.

3.2 Analysis of Individual Component Approaches

To create our ensembles, we followed the recommendations of Hansen and Salamon [25] in that we assembled and assessed a cross-discipline list of candidate ML, DL, and TL classifiers, focusing specifically on classifiers which had been applied to the task of sentiment analysis or emotion classification. In assessing



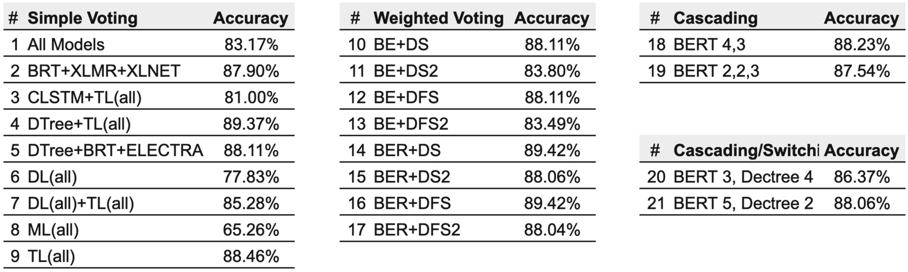
	Machine Learning Classifiers				Deep Learning Classifiers					Transformer Learning Classifiers				
Emotion	Decision Trees	Linear SVC	Logistic Regression	SVM	BiGRU	BiLSTM	C-LSTM	GRU	LSTM	BERT	ELECTRA	RoBERTa	XLM -RoBERTa	XLNet
joy	82.45%	79.35%	78.60%	82.01%	84.89%	85.15%	86.26%	79.33%	83.44%	90.86%	90.57%	90.33%	88.24%	89.92%
sadness	79.87%	64.48%	62.92%	41.76%	76.20%	78.99%	79.98%	71.70%	73.92%	89.50%	87.89%	87.86%	85.80%	87.42%
anger	83.71%	72.14%	68.71%	72.33%	82.45%	79.44%	83.20%	74.62%	82.00%	90.36%	89.89%	88.38%	87.20%	89.26%
love	77.79%	41.31%	35.98%	12.06%	66.24%	63.70%	68.11%	55.03%	60.96%	82.33%	80.53%	78.22%	76.13%	77.97%
thankfulness	81.64%	50.96%	46.46%	44.01%	69.06%	69.74%	71.30%	59.35%	64.47%	84.03%	82.39%	79.13%	77.64%	80.60%
fear	76.45%	28.38%	22.57%	8.45%	57.17%	55.25%	57.17%	42.84%	45.92%	76.94%	74.20%	70.40%	66.89%	73.24%
surprise	81.87%	9.66%	3.06%	1.78%	39.91%	42.61%	43.16%	18.38%	37.43%	64.49%	58.54%	55.22%	46.25%	55.28%

Fig. 1. Heatmap of classification accuracy by emotion for each classifier - greater than 80% - green, 50–80% - yellow, below 50% - red. (Color figure online)

individual models, we focused on base models and common implementations of each approach, including ML classifiers (decision trees, linear SVC, logistic regression, Naïve Bayes, SVM), DL classifiers (GRU, BiGRU, LSTM, C-LSTM, BiLSTM), and TL classifiers (BERT, ELECTRA, RoBERTA, XLM-R, XLNet). For additional detail on hyperparameter selection see [21]. We followed the same basic outline in assessing each model, in that we pre-processed our dataset and saved a clean version for reuse across all models compared. Next we trained or fine-tuned each model, performed 10-fold cross-validation to compute average accuracy, and created a heatmap (see Fig. 1) to assess how each performed in classifying specific emotions. This served to help identify strengths and weaknesses among individual component models, and informed the creation of the ensemble approaches we explored. We selected BERT, the most accurate singleton classifier, as a baseline for comparing ensemble performance.

3.3 Analysis of Ensemble Approaches

Based on the analysis of individual component approaches, we created 21 ensembles, including simple voting, weighted voting, cascading, and cascading/switching ensembles. Simple voting ensembles were created by pooling predictions from selected classifiers, as described by the names of their approaches (e.g. TL(all) is an ensemble including BERT, ELECTRA, RoBERTa, XLM-R, and XLNet), with each component receiving one vote per sample. Weighted voting ensembles were designed to leverage the greater accuracy of decision trees for the least represented classes in the HT dataset, adding votes from decision trees only when fear or surprise were predicted. The weighted voting ensembles are identified with abbreviations, where B is BERT, E is ELECTRA, R is RoBERTa, D is decision trees, F is fear, S is surprise, and 2 (when present) indicates that 2 votes were added whenever decision trees predicted fear (**BER+DS2**) or fear and suprise (**BER+DFS2**) instead of 1 vote. The cascading and cascading/switching ensembles were designed to append new super-class labels to the HT dataset to segment the data into subsets for training individual super-class and sub-class models. For example the cascading ensemble named **BERT 5, Dectree 2** indicates the super-classes were segmented to include the 5 most represented classes (joy, sadness, anger, love, and thankfulness) in one class and the 2 least represented (fear and surprise) in another class. A BERT model was trained to classify each sample as belonging to one of these super-classes and



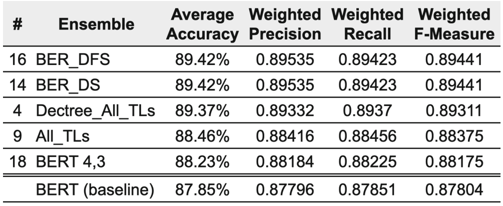
#	Simple Voting	Accuracy
1	All Models	83.17%
2	BRT+XLMR+XLNET	87.90%
3	CLSTM+TL(all)	81.00%
4	DTree+TL(all)	89.37%
5	DTree+BRT+ELECTRA	88.11%
6	DL(all)	77.83%
7	DL(all)+TL(all)	85.28%
8	ML(all)	65.26%
9	TL(all)	88.46%

#	Weighted Voting	Accuracy
10	BE+DS	88.11%
11	BE+DS2	83.80%
12	BE+DFS	88.11%
13	BE+DFS2	83.49%
14	BER+DS	89.42%
15	BER+DS2	88.06%
16	BER+DFS	89.42%
17	BER+DFS2	88.04%

#	Cascading	Accuracy
18	BERT 4,3	88.23%
19	BERT 2,2,3	87.54%

#	Cascading/Switchi	Accuracy
20	BERT 3, Dectree 4	86.37%
21	BERT 5, Dectree 2	88.06%

Fig. 2. Ensemble average accuracy.



#	Ensemble	Average Accuracy	Weighted Precision	Weighted Recall	Weighted F-Measure
16	BER_DFS	89.42%	0.89535	0.89423	0.89441
14	BER_DS	89.42%	0.89535	0.89423	0.89441
4	Dectree_All_TLs	89.37%	0.89332	0.8937	0.89311
9	All_TLs	88.46%	0.88416	0.88456	0.88375
18	BERT 4,3	88.23%	0.88184	0.88225	0.88175
	BERT (baseline)	87.85%	0.87796	0.87851	0.87804

Fig. 3. Comparing 5 most accurate ensembles with the BERT baseline.

this result was passed to one of two other models (a BERT model for the 5 most represented and a decision tree model trained to predict the 2 least represented classes) trained to predict from either the top 5 classes or bottom 2 classes respectively. The cascading hybrid **BERT 4,3** leverages one BERT model fine-tuned for the initial super-class prediction and two additional BERT models fine-tuned to predict within the sub-classes. The entire set of predictions was then reassembled and assessed for accuracy, with significance testing via ANOVA between the 5 most accurate models and the BERT baseline, as well as average accuracy, weighted precision, weighted recall, and weighted f-measure for each.

4 Results and Discussion

Of the individual classifiers we evaluated, the most accurate were the TL algorithms (in order from most to least accurate - BERT, ELECTRA, RoBERTa, XLNet, XLM-R), followed by decision trees, then all DL algorithms (C-LSTM, BiGRU, BiLSTM, LSTM, GRU in descending order), and finally the remaining ML algorithms (Linear SVC, Logistic regression, Naïve Bayes, SVM).

12 of 21 ensembles created were more accurate than the BERT baseline accuracy of 87.851%, including 4 of 9 simple voting ensembles, 6 of 8 weighted voting ensembles, 1 of 2 cascading ensembles, and 1 of 2 cascading/switching ensembles. The most accurate ensembles were weighted voting ensembles BER_DFS and BER_DS, with 89.423% average accuracy. Figure 2 shows accuracy across all tested ensembles and Fig. 3 shows a detail comparison of the accuracy, precision, recall, and f-measure for the top 5 ensembles and the BERT baseline. We also performed a single factor analysis of variance between BERT and the 5 most accurate ensembles and found that the variance was statistically significant, with

a p-value of 9.92e−59. The addition of weighted votes for fear appeared to have little affect on the accuracy of our ensembles, with no difference in accuracy scores for BER_DFS and BER_DS. The ensembles which were less accurate than the BERT baseline consisted primarily of reference models created to assess novel approaches rather than realistically expected to outperform the baseline.

5 Conclusions and Future Work

Results show that ensembles can provide more accurate results than the most accurate single classifier, with at least 5 ensembles providing significantly more accurate results than BERT (89.423% for our best ensemble compared to 87.851% for the baseline). These also showed performance improvement compared to the BERT baseline in precision, recall, and f-measure. Results also showed that simple voting, weighted voting, cascading, and cascading/switching ensembles may all provide measurably more accurate results, when designed to offset the weaknesses of one approach with the strengths of another approach.

Future work includes testing further ensemble variations, including dictionary classifiers, to understand tradeoffs in ensemble architectures, evaluation with additional EMDISM datasets under development, and extending our research to identify imbalance thresholds wherein voting and switching ensembles are most effective. Overall, results demonstrate the potential of ensemble approaches for performance improvement in EMDISM, with the potential to benefit a wide variety of applications that rely on accurate understanding of emotion contexts.

References

1. Al-Omari, H., et al.: EmoDet at SemEval-2019 task 3: emotion detection in text using deep learning. In: Proceedings of the 13th International Workshop on Semantic Evaluation (2019)
2. Araque, O., et al.: Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Syst. Appl. **77**, 236–246 (2017)
3. Asif, M., et al.: Sentiment analysis of extremism in social media from textual information. Telematics Inform. **48**, 101345 (2020)
4. Baziotis, C., et al.: Ntua-slp at semeval-2018 task 3: tracking ironic tweets using ensembles of word and character level attentive RNNs. arXiv:1804.06659 (2018)
5. Bickerstaffe, A., Zukerman, I.: A hierarchical classifier applied to multi-way sentiment detection. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 62–70. Association for Computational Linguistics (2010)
6. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
7. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adap. Inter. **12**(4), 331–370 (2002)
8. Burnap, P., et al.: Multi-class machine classification of suicide-related communication on twitter. Online Soc. Networks Media **2**, 32–44 (2017)
9. Cao, M.D., Zukerman, I.: Experimental evaluation of a lexicon-and corpus-based ensemble for multi-way sentiment analysis. In: Proceedings of the Australasian Language Technology Association Workshop 2012, pp. 52–60 (2012)
10. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)

11. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
12. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: pre-training text encoders as discriminators rather than generators (2020)
13. Da Silva, N.F., Hruschka, E.R., Hruschka, E.R., Jr.: Tweet sentiment analysis with classifier ensembles. Decis. Support Syst. **66**, 170–179 (2014)
14. De Choudhury, M., et al.: Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media (2013)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2019)
16. Duin, R.P.: Classifiers in almost empty spaces. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 2, pp. 1–7. IEEE (2000)
17. Duppada, V., Jain, R., Hiray, S.: Seernet at semeval-2018 task 1: domain adaptation for affect in tweets. arXiv preprint arXiv:1804.06137 (2018)
18. Efron, B.: The efficiency of logistic regression compared to normal discriminant analysis. J. Am. Stat. Assoc. **70**(352), 892–898 (1975)
19. Ekman, P.: Basic emotions. In: Handbook of Cognition and Emotion, pp. 45–60 (1999)
20. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
21. Frye, R.H., Wilson, D.C.: Comparative analysis of transformers to support fine-grained emotion detection in short-text data. In: The Thirty-Fifth International Flairs Conference (2022)
22. Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., Heck, L.: Contextual LSTM (CLSTM) models for large scale NLP tasks. arXiv preprint arXiv:1602.06291 (2016)
23. Gupta, N., Gilbert, M., Fabbrizio, G.D.: Emotion detection in email customer care. Comput. Intell. **29**(3), 489–505 (2013)
24. Gupta, S.: Applications of sentiment analysis in business. Towards Data Science. https://towardsdatascience.com/applications-of-sentiment-analysis-in-business-b7e660e3de69
25. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **10**, 993–1001 (1990)
26. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
27. Kang, M., Ahn, J., Lee, K.: Opinion mining using ensemble text hidden Markov models for text classification. Expert Syst. Appl. **94**, 218–227 (2018)
28. Khan, J.: Sentiment analysis : Key to empathetic customer service. Ameyo. https://www.ameyo.com/blog/sentiment-analysis-key-to-empathetic-customer-service
29. Lample, G., Conneau, A.: Cross-lingual language model pretraining (2019)
30. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision. LNCS, vol. 1681, pp. 319–345. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-46805-6_19
31. Liu, Y., et al.: Roberta: a robustly optimized Bert pretraining approach (2019)
32. Oussous, A., Lahcen, A.A., Belfkih, S.: Impact of text pre-processing and ensemble learning on Arabic sentiment analysis. In: Proceedings of the 2nd International Conference on Networking, Information Systems & Security, p. 65. ACM (2019)
33. Perikos, I., Hatzilygeroudis, I.: Recognizing emotions in text using ensemble of classifiers. Eng. Appl. Artif. Intell. **51**, 191–201 (2016)

34. Pujol, F.A., Mora, H., Pertegal, M.L.: A soft computing approach to violence detection in social media for smart cities. Soft. Comput. **24**(15), 11007–11017 (2019). https://doi.org/10.1007/s00500-019-04310-x
35. Ramadhan, W., Novianty, S.A., Setianingsih, S.C.: Sentiment analysis using multinomial logistic regression. In: 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC), pp. 46–49. IEEE (2017)
36. Ranganathan, J., Hedge, N., Irudayaraj, A., Tzacheva, A.: Automatic detection of emotions in twitter data-a scalable decision tree classification method. In: Proceedings of the RevOpID 2018 Workshop on Opinion Mining, Summarization and Diversification in 29th ACM Conference on Hypertext and Social Media (2018)
37. Schapire, R.E.: The strength of weak learnability. Mach. Learn. **5**(2), 197–227 (1990)
38. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
39. Smetanin, S.: Emosense at semeval-2019 task 3: Bidirectional LSTM network for contextual emotion detection in textual conversations. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 210–214 (2019)
40. Symeonidis, S., et al.: A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Syst. Appl. **110**, 298–310 (2018)
41. Vapnik, V.: The nature of statistical learning theory. Springer, New York (2000). https://doi.org/10.1007/978-1-4757-3264-1
42. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)
43. Walther, C.: Sentiment analysis in marketing: What are you waiting for? CMS Wire. https://www.cmswire.com/digital-marketing/sentiment-analysis-in-marketing-what-are-you-waiting-for/
44. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: a multitask benchmark and analysis platform for natural language understanding (2019)
45. Wang, W., et al.: Harnessing twitter "big data" for automatic emotion identification. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 587–592. IEEE (2012)
46. Wang, X., et al.: A novel hybrid mobile malware detection system integrating anomaly detection with misuse detection. In: Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services, pp. 15–22. ACM (2015)
47. Wolfe, J.: Want faster airline customer service? try tweeting. The New York Times. https://www.nytimes.com/2018/11/20/travel/airline-customer-service-twitter.html
48. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci. **181**(6), 1138–1152 (2011)
49. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding (2020)
50. Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. Knowl. Inf. Syst. **60**(2), 617–663 (2018). https://doi.org/10.1007/s10115-018-1236-4
51. Yue, T., Chen, C., Zhang, S., Lin, H., Yang, L.: Ensemble of neural networks with sentiment words translation for code-switching emotion detection. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2018. LNCS (LNAI), vol. 11109, pp. 411–419. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99501-4_37
52. Zhang, L., et al.: Deep learning for sentiment analysis: a survey. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **8**(4), e1253 (2018)