Detecting Fake News Through Emotion Analysis

Andrew L. Mackey
Computer Science and Engineering
University of Arkansas
Fayetteville, Arkansas, USA
almackey@uark.edu

Susan Gauch
Computer Science and Engineering
University of Arkansas
Fayetteville, Arkansas, USA
sgauch@uark.edu

Kevin Labille
Computer Science and Engineering
University of Arkansas
Fayetteville, Arkansas, USA
kclabill@uark.edu

***Abstract*—Automating the detection of fake news is a challenging problem for the research community due to the various degrees of falsified information and ways in which it can be classified. In this work, we present a Bidirectional Encoder Representations (BERT)-based machine learning model that captures linguistic and emotional features of a document to improve the task of classifying misinformation. The different types of psychological emotions are presented along with the methods used to capture the emotions of words. We investigate how different emotional features can augment existing data to facilitate the detection of fake news and improve upon existing baseline results. Our work demonstrates the ability for emotional features, when combined with other word-embedding models, such as BERT, to improve the performance benchmarks of fake news detection tasks.**

Index Terms*—*Fake news classification; misinformation; emotion analysis natural language processing

I. Introduction

Social media providers and other distributors of online content are facing increasing pressure to find ways to curtail the spread of falsified information with the intention to deceive users, while also balancing the legalities and potential repercussions from actions taken. Furthermore, journalists who author publications that run contrary to the primary views held by certain groups find themselves being labeled as fake or misleading, even in situations where content was authored solely for the purpose of entertainment. This compels organizations responsible for managing content to differentiate between information as being factually true, misleading, factually untrue for the purpose of entertainment, or blatantly false with the intent to deceive others, often for malicious purposes.

Organizations have been established with the purpose of investigating content and measuring the accuracy of various claims. In recent years, the number of companies dedicated to this task has increased [27]. For example, PolitiFact analyzes comments that were made and ranks them on a scale with values between between true and false, rather than strictly true or false. Small claims are analyzed for the degree of their truthfulness. The task of evaluating the degree of truthfulness is challenging as individuals can easily misidentify claims as being true despite small discrepancies in the way a claim was worded.

Fake news detection using emotion analysis is a classification problem, either binary or multi-class, involving the creation of emotion vectors to augment lexical features and machine learning algorithms to effectively identify content that contains misinformation. Emotion analysis involves the utilization of techniques, mostly derived from lexicons and machine learning algorithms, to extract the psychological associations between words and emotions. Research experiments have been conducted using artificial intelligence and machine learning algorithms to identify and detect falsified content [27]. While there have been significant advancements in the fields of machine learning and natural language processing to tackle and identify fake news articles, additional work is necessary to improve our ability to handle different types of fake news effectively [33]. The identification of smaller text claims, such as social media posts, have not received the same amount of coverage as other forms of fake news (i.e. propaganda, falsified news articles, etc.). Similarly, different types of fake information, such as satirical publications, may receive incorrect classifications in spite of the fact that no malicious intent was assumed.

In this paper, we present an analytic study covering the emotional content contained within varying types of news articles. A model is introduced for incorporating emotion analysis in fake news detection tasks to mitigate the spread of misinformation intending to deceive users. In doing so, the efficiency of emotion vectors are demonstrated as a way to improve existing models. Furthermore, we propose a neural network model for incorporating emotion analysis with word embedding vectors produced by through the Bidirectional Encoder Representations (BERT) model.

II. Related Work

In the following sections, we consider previous work in the areas of emotion analysis and fake news detection.

A. Emotion Analysis

Numerous fields addressing affective computing [16] have demonstrated an interest in the study of emotions and the implications it has for human-computer interaction. The emotion analysis of text allows for the latent emotions and sentiment of words, phrases, and sentences to be extracted. Emotion analysis is often analogous to applications of opinion mining and sentiment analysis [14] and the study of affective lexicons from the field of pyscholinguistics, which evaluates the relationship between psychological processes and linguistic

 ISBN: 978-1-61208-874-7

behaviors [4]. In contrast to opinion mining and sentiment analysis where polarity is often measured, emotion analysis aims to associate text with a predefined set of psychological models as determined by the dimensions of valence, activation, and control [19] [22] [23].

Prior studies in the field of psychology focused on the universality of emotions [7] [9] [10]. Six emotions were originally emphasized as being *universal*: ANGER, DISGUST, FEAR, JOY, SADNESS, and SURPRISE [11]. In general, these emotions are represented in models as a discrete set of possibilities or as a domain-general scale (valence, arousal, etc.). Debate over the topic of emotion models still persists in research literature with some researchers proposing categories that are highly dimensional [6] and others suggesting emotions are organized along affective dimensions [2]. Studies questioned the qualitative differences between emotions [26] and the possibility of an existence of overlapping affective features between emotion categories [2].

Emotion classification is typically categorized as being 1) rule-based or 2) machine learning. In earlier implementations of rule-based techniques, authors build or expand existing lexicons of varying emotional characteristics to identify words in data sets evoking emotional features. Techniques for annotating these lexicons involve either crowdsourcing or curation by experts. One study was conducted to model the independent, neurophysiological systems of valence and arousal of social media posts to produce a data set and model that measures the affective norms of subjective social media postings [17]; this served as a departure from prior work that focused predominantly on valence or sentiment [24] [25]. The model proposed utilized the circumplex model of affect with emotions being projected into a vector space of valence, arousal, and dominance [19]. Another study evaluated the concreteness and abstractness of social networking data while measuring emotional intensity [12].

Several advancements in the fields of machine learning and natural language processing have paved the way for new methods of learning semantic relationships between words and emotions. The goal these algorithms is to improve upon dictionary techniques by utilizing supervised machine learning algorithms over lexical features, such as n-grams, word embeddings, and affect lexicons [1]. Machine learning techniques are then able to categorize and predict the appropriate emotion category for text. Many state-of-the-art methods utilize pre-trained word embeddings to extract features using unsupervised machine learning [1] [5] [13] [15] . Through these embeddings, words can be projected into a space such that they are represented as function of their context words.

B. Fake News Detection

Fake news is defined broadly as being news articles that demonstrate the intention of being verifiably false to mislead consumers of this information for entertainment or deceptive purposes. Fake news, while not necessarily a new topic, is one that has received considerable attention from both the public and academic research communities. Similar to other terms that are loosely defined, fake news has many varying definitions between authors and publications. Consider the situation of satirical publications. Whereas some authors include these types of articles as fake news, other authors narrow the definition to news articles as fabrications, hoaxes, or news that is, otherwise, deliberately false with negative intentions, despite attempts to convey the entertainment goal of the articles.

As the aim of each type of fake news differs, we define each of the following as the distinct categories used for the classification of fake news: satire, hoax, propaganda, and clickbait. *Satire* represents a collection of articles where the author of the article intends to entertain the reader through misinformation, sarcasm, or fabrications [27]. It is important to note that an author of satirical work does not intend to mislead the reader. Unlike satire, *hoaxes* are false articles passed as truth, often with the intent of humorous deception. *Propaganda* are articles that are false and meant to deliberately harm a specific party. *Clickbait* is a type of article where the goal is to obtain a reader's attention through misleading headlines, images, etc. that do not align with the perceived goal of the article. It is important to stress that the underlying motivation of the work to deceive, as demonstrated in satire, is a component used in distinguishing the type of fake news a document is classified as being.

The advent of social media, accompanied by the widespread adoption of these services, has proven to be problematic for news consumption by users. Information is able to flow through these social networks rapidly in a manner that is cheap and easy to access. With few limitations in place, it enables the dissemination of fake, misleading, or erroneous news through these same networks, often unabated [28]. Consequently, deceptive practices of misleading or shifting public opinion in a particular direction could adversely influence groups of individuals in social networks based on false pretenses. Fake news increases the mistrust individuals have in real news as users express more skepticism in all information.

In general, there are three characteristics demonstrated in prior work for fake news detection [18]: the content of the article, response from users as a result of posting the article, and the source of the article. Automating the detection of fake news is challenging for several reasons. A number of studies demonstrate the difficulty users have in discerning whether or not an article is fake [3] [8] . The intentionally misleading nature of fake news curtails attempts to categorize documents as being real and fake by the content alone [21]. This presents numerous challenges unique to this task [20] [21]. Variations of the original content is often spread through social media, thus exacerbating the problem of classifying fake news while adding additional complexity due to the additional noise. Prior work has demonstrated that auxiliary information is needed to facilitate the classification of news.

C. Dataset

The data used for this was the publicly available dataset from [27], which is comprised of news articles obtained from crawling seven different unreliable news sites, including The Onion, The Borowitz Report, Clickhole, American News, DC Gazette, The Natural News, and Activist Report.. The types of news were defined as being *satire*, *hoax*, *propaganda*, or *trusted*. For the trusted news source, we include data from [29] where the authors constructed an approach to building a supervised reading comprehension dataset with news articles obtained from convolutional neural networks ($n = 90,266$). We limited the number of documents from the CNN dataset to a randomly extracted sample of $n = 10,000$ documents to limit the overrepresentation of any specific class.

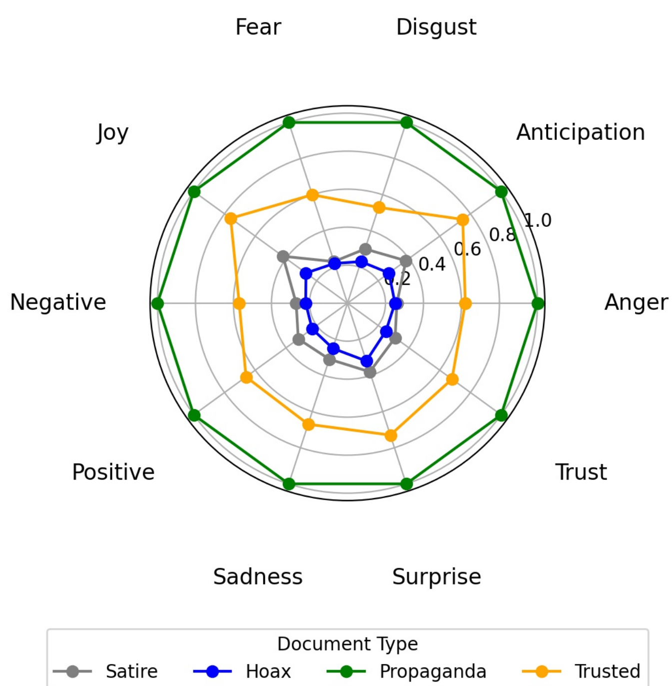

Fig. 1. The emotion feature frequency is shown for each document type, normalized by the maximum of each category.

Table II summarizes the type of news articles, document frequencies, mean document lengths and standard deviations, and median document lengths. Figure 1 visually represents this data normalized by the maximum for each category. New articles from the propaganda class have a higher average number of tokens than other classes. When considering the robustness of the statistical measures to control for outliers, the median of the propaganda class is marginally higher than the trusted class. All data is preprocessed using standard natural language preprocessing techniques, including down-casing, stopword removal, tokenization, etc. We utilize the NLTK toolkit for computational linguistic analysis. The overall distribution of the data can be seen in Figure 2.

III. Fake News Detection

The representation of sentiment as a set of psycholinguistic features has been of interest in prior literature in the field of natural language processing. We augment this work by conducting several experiments to determine which combinations of feature sets yield the best predictive capabilities for the classification of fake news. Our goal is to demonstrate the efficiency of emotion vectors and prove the efficacy of augmenting existing feature sets with emotion features to facilitate classification tasks. To this end, we construct three baseline models for automated fake news detection and compare several models that leverage these emotion vectors. The models, parameters, and configurations are described in the following sections. The models are evaluated using the datasets as described below.

A. Overview

In our experimentation tasks, we evaluated multiple classification algorithms – support vector machines, logistic regression, etc. – and found neural network models to perform the best with a word embedding features. Each document is represented in the training set as a vector of size $n = |V|$ where V is the lexicon derived from the training data. The second baseline model constructed uses the word embeddings formed by extracting fixed-length feature representations from the words in a variable-length documents [30].

B. Emotion Vectors

Our model $E = \{E_1, E_2, E_3\}$ leverages emotions from the discrete and continuous sets of the following emotions and sentiment. We define the set of emotions E with the following ($|E| = 12$):

$$E_1 = \{ \textit{anger}, \textit{anticipation}, \textit{disgust}, \textit{fear}, \textit{joy}, \textit{sadness}, \textit{surprise}, \textit{trust} \}$$
$$E_2 = \{ \textit{positive}, \textit{negative} \}$$
$$E_3 = \{ \textit{valence}, \textit{arousal} \}$$

For each document, an emotion vector is generated the aggregation of tagged words in the EmoLex emotional resource [31]. We investigated variations of the vector as seen in Table 2. The first approach, EMO$_{SUM}$, is an emotion vector produced by aggregating the sum of each emotion for each word tagged in the document. EMO$_{ZS}$ represents the vector of z-scores for each emotion, such that every emotion e_i is calculated as:

$$z_i = \frac{e_i - \bar{e}_i}{\sigma_{e_i}}$$

Finally, we consider normalizing the vectors using a relative maximum EMO$_{RM}$ for each emotion feature e_i as:

$$RM(e_i) = \frac{e_i}{\underset{d \in D}{\arg\max}(e_i)}$$

Our next task was to determine how to incorporate the number of matching tokens with the emotion scores produced. EMO$_{RM1}$ represents relative maximum of the emotion scores multiplied by the number of matching tokens, whereas EMO$_{RM2}$ is the relative maximum of the emotion scores divided by the number of matching tokens. After testing

TABLE I
NEWS ARTICLES AND MEAN EMOTION TOKENS PER DOCUMENT AND STANDARD DEVIATION

Type	**Anger**	**Anticipation**	**Disgust**	**Fear**	**Joy**	**Negative**	**Positive**	**Sadness**	**Surprise**	**Trust**
Satire	5.3 ± 6.8	8.8 ± 8.5	3.5 ± 4.6	7.2 ± 9.3	6.4 ± 7.0	11.6 ± 12.5	18.5 ± 17.2	5.5 ± 6.6	3.8 ± 4.2	12.0 ± 11.8
Hoax	5.1 ± 5.4	6.2 ± 5.8	2.7 ± 3.2	6.9 ± 7.6	4.1 ± 4.8	9.5 ± 9.0	13.2 ± 12.1	4.4 ± 5.0	3.2 ± 3.3	9.6 ± 8.7
Prop.	20.5 ± 33.5	23.1 ± 31.3	12.0 ± 23.7	31.1 ± 48.6	15.2 ± 22.3	44.0 ± 71.8	57.0 ± 74.7	18.0 ± 31.4	10.0 ± 15.7	38.9 ± 50.0
Trusted	12.7 ± 11.0	17.3 ± 11.2	6.3 ± 6.2	19.0 ± 15.1	11.5 ± 9.4	25.1 ± 17.4	37.6 ± 22.0	12.1 ± 9.6	7.3 ± 5.5	26.3 ± 16.0

TABLE II
NEWS ARTICLES WITH NUMBER OF DOCUMENTS, AVERAGE DOCUMENT LENGTHS, AND MEDIAN DOCUMENT LENGTHS

Doc. Type	**# of Docs**	**Avg. Tokens**	**Med. Tokens**
Satire	$13,942$	206 ± 177	105
Hoax	$6,892$	141 ± 122	109
Propaganda	$15,061$	587 ± 808	458
Trusted	$9,681$	428 ± 205	401

both techniques, we constructed each emotion vector for its corresponding document using word frequencies normalized by the number of matching tokens (EMO$_{RM2}$).

C. Baseline Models

We investigate the impact of both the emotion and extended emotion feature vectors due to their efficiency for the fake news detection task . The first model is constructed by utilizing emotion features obtained from the input documents. We construct a feed forward neural network architecture with two fully connected layers with 512 neurons using the rectified linear unit (ReLU) activation function. Following each fully connected layer, we implement a Dropout layer with dropout rates of 0.5 and 0.3, respectively. We add a final dense layer as the output using the softmax activation function with the number of units corresponding to the number of $\hat{\mathbf{y}}$ target classes. In previous sections, we introduced the methods by which we encode news articles and construct emotion vectors for each article. We define $\hat{\mathbf{y}}$ as the predicted probability of the target class being fake or real news. The procedure would be similar in a multi-class classification problem for detecting hoaxes, propganda, clickbait, satire, or legitimate news. We define $\mathbf{d}$ and $\mathbf{e}$ as the learned features for news documents and emotion vectors, respectively. Furthermore, $\mathbf{b}$ is defined as the bias term and $\mathbf{W}$ represents the learned weights.

$$\hat{\mathbf{y}} = \mathrm{softmax}([\hat{\mathbf{d}}, \hat{\mathrm{e}}]\mathbf{W} + \mathbf{b})$$

The batch size was set to 64 and we implemented early stopping criteria to limit potential overfitting. We utilize the Adam optimization algorithm and a categorical cross-entropy loss function for this multi-class classification task. A learning rate of 0.001 was used.

The second model is constructed by forming document-level word embeddings from BERT for each of the input documents [32]. The BERT embeddings were formed from $L = 12$ hidden layers (transformer blocks), with a hidden size of $H = 128$ and $A = 2$ attention heads. After the BERT layers, we implement the same feed forward network architecture as described above. The final model architecture was formed by using bag of words feature vectors using TF-IDF weights. To measure the impact and effectiveness of emotion vectors, we consider the top k features for the BOW model. We established $k = 128$ for comparison to the BERT model. The feature vectors were normalized using min-max scale.

All documents containing a low number of tokens or convey no emotional content such that the magnitude of the vector $\|\mathbf{e}\| = 0$ were removed from the document corpus. Each experiment was conducted from training, testing, and validation splits of sizes 0.7, 0.2, and 0.1, respectively. The mean performance metric from each experiment conducted 10 times from random shuffles of the data is reported.

IV. EVALUATION

Having presented models for the task of identifying fake news, we evaluate the models using the data described in earlier sections. Our hypothesis is that emotion vectors can improve the detection of fake news detection by augmenting existing models with additional information. Given the complexity of fake news detection, we expect that emotion analysis alone may not be suitable to compress the information needed correctly identify falsified information. The experiments are therefore designed to evaluate the effectiveness of emotion analysis in the classification of fake news. First, we want to establish whether or not emotion features can be used in fake news detection. Second, we compare our baseline models to those where features have been augmented with emotions. Third, we want to measure the efficiency of emotion features by evaluating the gains achieved through adding emotional

 ISBN: 978-1-61208-874-7

TABLE III
FAKE NEWS CLASSIFICATION METHODS WITH EACH OF THE PROPOSED MODELS

Type	Method	Accuracy	Precision	Recall	F1
BASELINE	EMO+NN	0.569	0.692	0.308	0.423
	BERT+NN	0.763	0.798	0.721	0.757
WORD EMBED	EMOEX+NN	0.593	0.704	0.369	0.482
	EMO+BERT+NN	0.792	0.824	0.753	0.786
	EMOEX+BERT+NN	0.794	0.823	0.754	0.786
BAG OF WORDS	BOW+NN	0.793	0.795	0.792	0.794
	BOWEX+NN	0.798	0.799	0.797	0.798
	EMO+BOW+NN	0.861	0.863	0.857	0.861

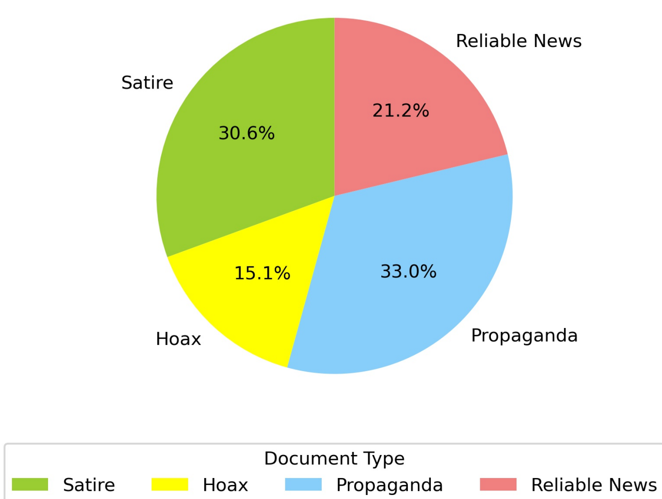

Fig. 2. The distribution of the dataset used for classification is presented by the news article type.

context to existing models in comparison to other lexical feature additions.

TABLE IV
BASELINE MODEL EVALUATION

Method	Accuracy	Precision	Recall	F1
EMO+NN	0.569	0.692	0.308	0.423
EMOEX+NN	0.593	0.704	0.369	0.482
BERT+NN	0.763	0.798	0.721	0.757

The results presented in Table III demonstrate that emotional features can enhance existing models to improve the classification of fake news. For our baseline models as seen in Table IV, we consider the emotional features baseline EMO+NN or word embeddings baseline BERT+NN produced from BERT to use for training. The EMO+NN model using emotion vectors reported a baseline accuracy of 0.569, whereas the BERT+NN word embeddings model produced 0.763 for the same task. For our experimental models, we augment the existing feature vectors with our emotion vectors EMO+BERT+NN or the extended emotion vector EMOEX+BERT+NN for the document. The concatenation of emotion vectors and BERT word embeddings for neural network classifier improved the accuracy and F1 metrics by 2.9%. This can be observed in Table V. Similarly, the extended emotion vectors EMOEX+BERT+NN improved the accuracy performance by 3.1% and the F1 score by 2.9%. When considered individually, EMOEX+BERT+NN had an overall accuracy improvement over the baseline EMO+BERT+NN by 2.4% and 5.9% for the F1 score.

TABLE V
EVALUATION OF WORD EMBEDDINGS AND EMOTION FEATURE MODELS

Method	Accuracy	Precision	Recall	F1
BERT+NN	0.763	0.798	0.721	0.757
EMO+BERT+NN	0.792	**0.824**	0.753	0.785
EMOEX+BERT+NN	**0.794**	0.823	**0.754**	**0.786**

To demonstrate the application of emotion vectors to other tasks, we consider a model trained with TF-IDF weighted bag of words feature vectors BOW with $k = 128$ features. The model achieves an accuracy performance of 0.793 and F1 score of 0.794. The impact from adding the emotion vectors to the model is demonstrated in EMO+BOW+NN. The model achieves an accuracy and F1 score of 0.861, which is a 6.8% improvement over the BOW model for comparison. Similarly, we consider the impact of adding $|\text{EMO}| = 18$ features to the $k = 128$ top features selected for the BOW model. An additional 18 features are added to the top k features to produce an expanded bag of words feature vector of length $k = 146$ to produce model BOWEX+NN. By increasing the

BOW model by the same number of features as the emotion lexicon, we obtain an accuracy of 0.798, which is a marginal improvement of 0.5%. The improvements from increasing the feature vectors with additional word features did not have the same measurable impact as adding the same number of emotion features as seen in Table VI.

TABLE VI
COMPARISON OF BAG OF WORDS MODELS AND EMOTION FEATURES

Method	Accuracy	Precision	Recall	F1
BOW+NN	0.793	0.795	0.792	0.794
BOWEX+NN	0.798	0.799	0.797	0.798
EMO+BOW+NN	0.861	0.863	0.857	0.861

The experiments presented here demonstrate the ability of emotion features to facilitate the classification of fake news in a multiclass environment. The stronger improvements to bag of words models over word embedding models suggests that word embeddings capture additional semantics in lexical meanings that are otherwise not present in bag of words models. Gains were similarly observed in word embedding models, and subsequently demonstrating the ability for emotion features to improve existing models.

V. CONCLUSION

The utilization of emotion analysis and features for improving existing machine learning tasks in the detection of fake news provides a promising track for building systems capable of understanding the patterns of information with intentions of deceiving the user. The effectiveness of applying emotion features to fake news detection and existing frameworks or models was demonstrated. The incorporation of other sources of data into models may be necessary to expand beyond the tasks described here. Given the ability of emotions to distinguish between targets in a multiclass setting, further experimentation will need to be conducted to better understand how to improve upon existing techniques for extracting emotional context through a combination of lexicon and machine-learning based techniques.

REFERENCES

[1] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. *arXiv preprint arXiv:1709.04219*, 2017.

[2] Lisa Feldman Barrett, Zulqarnain Khan, Jennifer Dy, and Dana Brooks. Nature of emotion categories: Comment on cowen and keltner. *Trends in cognitive sciences*, 22(2):97–99, 2018.

[3] Michael Barthel, Amy Mitchel, and Jesse Holcomb. Many americans believe fake news is sowing confusion. *Pew Research Center*, 2016.

[4] Gerald L Clore, Andrew Ortony, and Mark A Foss. The psychological foundations of the affective lexicon. *Journal of personality and social psychology*, 53(4):751, 1987.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[6] Alan S Cowen and Dacher Keltner. Clarifying the conceptualization, dimensionality, and structure of emotion: Response to barrett and colleagues. *Trends in cognitive sciences*, 22(4):274–276, 2018.

[7] Charles Darwin. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1872.

[8] Camila Domonoske. Students have 'dismaying' inability to tell fake news from real, study finds. *National Public Radio*, 23, 2016.

[9] Paul Eckman. Universal and cultural differences in facial expression of emotion. In *Nebraska symposium on motivation*, volume 19, pages 207–284, 1972.

[10] Paul Ekman. *Darwin and facial expression: A century of research in review*. Ishk, 1973.

[11] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[12] Maximilian Köper, Evgeny Kim, and Roman Klinger. IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

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

[14] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135, January 2008.

[15] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[16] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[17] Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 9–15, 2016.

[18] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[19] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[20] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, 2019.

[21] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[22] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, 2008.

[23] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[24] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

[25] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354, 2005.

[26] Christine D Wilson-Mendenhall, Lisa Feldman Barrett, and Lawrence W Barsalou. Variety in emotional life: within-category typicality of emotional experiences is associated with neural activity in large-scale brain networks. *Social cognitive and affective neuroscience*, 10(1):62–71, 2015.

[27] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: analyzing language in fake news and political fact-checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, 2017.

[28] Castillo, Carlos and Mendoza, Marcelo and Poblete, Barbara Information credibility on twitter *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.
[29] Karl Moritz Hermann, Tomáš Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. 2015.
[30] Quoc V. Le and Tomas Mikolov Distributed representations of sentences and documents. 2014.
[31] Mohammad, Saif M. and Turney, Peter D. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 2013.
[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* 2018.
[33] Xinyi Zhou and Reza Zafarani A survey of fake news: fundamental theories, detection methods, and opportunities *ACM Comput. Surv* 2020.