Linguistic Features in Automatic Sarcasm Detection A Comparsion Across Two Social Media Platforms

Lauren Kruse

Montclair State University
Montclair, NJ
{krusel1@montclair.edu}

Abstract

This paper explores various linguistic features that contribute to sarcasm detection across two social media platforms, Reddit and Twitter. The linguistic features that are investigated are a combination of text and word count, stylistic, and psychological features. The features are first run on a corpus of tweets and then separately on a corpus of Reddit threads. Both sets of data are experimented on with and without context to test the importance contextual information plays in computational sarcasm detection. The two results of each social media platform are compared. The results of the experiment indicate that contextual information is essential for sarcasm prediction in both Reddit and Twitter. One key observation is that the presence of hashtags are an indicator of sarcasm detection. Lastly, sarcastic responses are typically associated with a negative emotion; anger, and will also respond with the same negative emotion. In contrary, non-sarcastic texts generally will reply to a negative emotion with a positive one; joy.

1 Introduction

Sarcasm is a figurative language device used by a speaker or writer to convey the opposite meaning of what they are actually saying. Sarcasm can also be closely mistaken with verbal irony, however, sarcasm differentiates itself by having the negative intent attached with it. Sarcasm typically aims to mock, or convey a negative emotion towards the subject. In verbal communication, body language, pause,or intonation can provide enough information to detect whether there is a presence or absence of sarcasm. However, the issues arise in writing, due to the fact the cues are inaccessible and not available to the reader. Without these cues, the reader must rely fully on the understanding of the world, the writer, and the surrounding context of the statement to determine if the assertion is sarcastic or a genuine utterance. The task of detecting sarcasm in speech has proven to be so difficult and subjective to the reader that social media users moderate their own comments using symbols and hashtags such as /s and #sarcasm to denote the sentiment on Reddit and Twitter, respectively. Matter-of-factly, the data sets that were used in this experiment were collected using those hashtags and denotations. (Ghosh et al., 2020).

In regards to computational methods of computing sarcasm detection, machines lack the real-word knowledge that is detrimental to their understanding of sarcasm. This limits their ability to classify whether or not the intent is sarcastic or non-sarcastic accurately. Many aspects of natural language processing are not able to be utilized to their best ability due to the lack of knowledge computers have. Beyond social-media conversations, assessing product reviews as positive or negative requires an understanding of both rhetorical and literary devices. Back in 2012, BIC rolled out a "For Her" line of pens which led their intended female audience to poke fun at the misogynist message of the product. One reviewer commented, "Well at last pens for us ladies to use. . . now all we need is "for her" paper and I can finally learn to write!". While this review seems positive and gave the product four stars, our understanding of the social climate today leads us to conclude that this review is sarcastic and should be classified as such.

In social media forums, new colloquialisms are constantly being introduced, along with a library of new emojis. Often these new colloquialisms and emojis are used by the writers to negate the sentiment of the text. Stylistic devices and features are also utilized to covey the opposite meaning from its original text. To obtain very high and effective results, deep learning models and transformer-based-architectures have been used to aide in sarcasm

detection. However, these models provide a "black box approach". This approach gives very little insight as what to features are present in automatic sarcasm detection. The sole purpose of this research work is to gain knowledge and learn the linguistic patterns that are associated with sarcasm detection. Because sarcasm is more than just a positive or negative emotion like most sentiment analyzers and the borders of sarcasm are not well defined, the goal is to investigate which linguistic features play the largest role. Furthermore, investigate if these top features are enough to detect sarcasm. Because the boarders of sarcasm are not well defined, the linguistic hypothesis of this paper is that all the features run simultaneously through the classifier will yield in the most promising and highest results. Lastly, the results of two social media platforms are compared and contrasted with each other. While this paper only researches two social media platforms, further work would divulge into numerous social media forums to test the results of the features and classifier. The features would also be tested with the utilization of a transformer-based-architecture.

2 Previous Work

The field of automatic sarcasm recognition has become extremely active in recent years. The most current event is the shared task (Ghosh et al., 2020) organized as a part of the 2nd Figurative Language workshop at ACL 2020. The task is typically framed as a binary classification task (sarcastic vs. non-sarcastic) considering either an utterance in isolation or in combination with contextual information. Early approaches to automatic sarcasm detection rely on different types of features, including sarcasm detonations, word embeddings, emojis, patterns between positive and negative emotions (e.g., Davidov et al. 2010; Tsur et al. 2010; González-Ibáñez et al. 2011; Riloff et al. 2013; Maynard and Greenwood 2014; Wallace et al. 2015; Ghosh et al. 2015; Joshi et al. 2015; Veale and Hao 2010; Liebrecht et al. 2013). Buschmeier et al. (2014) explore a range of features, mainly focused on sentiment, for the detection of verbal irony in product reviews. While this paper provides a good baseline for irony classification, the data in this paper differs in that it includes conversational threads between more than one speaker with many layers of contextual information.

More recent approaches apply deep learning methods (e.g., Ghosh and Veale 2016; Tay et al. 2018; Wallace et al. 2015). There is a great amount of research exploring the role of contextual information for sarcasm detection (e.g., Joshi et al. 2015; Bamman and Smith 2015; Misra and Arora 2019; Bamman and Smith 2015; Khattri et al. 2015; Amir et al. 2016; Rajadesingan et al. 2015; Ghosh and Veale 2017; Schifanella et al. 2016; Cai et al. 2019; Castro et al. 2019). Ghosh et al. (2020) report that almost all systems submitted as part of the shared task have used the transformer architecture, such as BERT (Turc et al. 2019) or RoBERTa (Liu et al. 2020), and other variants. They performed better than RNN architectures, even without any task specific fine-tuning. Unfortunately, it is difficult to explore what these transformers capture in regards to sarcastic tweets/threads and their context. The approach this paper takes uses a standard supervised classification model to gain insight on which natural language features help classify sarcasm on social media platforms. In this paper, linguistic features are categorized into three broad groups, experiments are run utilizing different combinations, taking a closer look at the role contextual information plays, and lastly, a comparison of how the features perform over two social media accounts.

3 Approach

The approach this paper takes utilizes a combination of count, stylometric, and psychological linguistic features to automatically detect the presence or absence of sarcasm in a given text. This approach intentionally experiments with a classical machine learning algorithm to get a better understanding of the patterns of linguistic features that contribute to sarcasm detection. The linguistic hypothesis is that there will be a difference between the linguistic features corresponding with the responses and contexts labeled as sarcastic. Sarcastic tweets and threads are likely to be semantically or emotionally in-congruent with their preceding tweets and threads, while non-sarcastic tweets/threads show a greater similarity with their context. To measure the emotional importance of a response and its context, a number of sentiment and emotion related features are extracted. The emotions explore a deeper layer what the writer is trying to convey, rather than just looking at the words in a given text. The emotions are then looked at over both classes and compared. The classifier is then used to test the importance of the features by

considering just the response, versus the response with its contextual counterpart. The objective is to test if the top features are selective enough to detect sarcasm, or utilizing all the features in conjunction with each other results in the most promising outcome. Lastly, each social media platform is run through the classifier separately and the results are compared and contrasted.

4 Data Set

Both corpora; Twitter and Reddit, were extracted from from the CodaLab shared task on sarcasm detection (Ghosh et al., 2020). The training data consists of 2,500 tweets labeled 'SARCASM' and 2,500 tweets labeled 'NON SARCASM', the balanced test data consists of an additional 1,800 labeled tweets. Reddit's balanced training data consisted of 4,400 labeled threads, while having the same balanced data set as Twitter of 1,800. Ghosh et al. (2020), this is a self-labeled data set where the tweets and threads are annotated as sarcastic based on the hashtags used by the users. The non-sarcastic tweets are the ones that do not contain the sarcasm hashtags, but may be labeled with either positive or negative sentiment hashtags, such as '#happy'. Retweets, duplicates, quotes, etc., have been excluded from the data set. Visit Ghosh et al. 2020 for more information on the shared task and data set. Each sarcastic and non-sarcastic tweet/thread is accompanied with an hierarchical conversation thread, e.g., context/1 is the immediate context, context/0 is the context that preceded context/1, and so on. The training and test data include up to 19 preceding tweets/threads labeled as context/0, context/1, . . . , context/19 (if available). For the purpose of this paper, only the response, context/0 and context/1 have been included in the research.

Platform	Lines	Words	Characters	Hashtags
Twitter	1,800	153,047	811,086	2,332
Reddit	1,800	101,461	570,987	36

Table 1: Data Comparison on Testing Data Set

5 Feature Extraction

This research predominately focuses on the linguistic patterns and the role they play in sarcasm detection. The features have been classified into three broad categories: *count, stylistic, and psychological*. Abonizio et al. (2020) defines count features as linguistic features that capture the overall objective of the context at the word and sentence level. The features utilized under the category of count features include; word-level count vectors, word-level tf-idf, n-gram word-level tf-idf, n-gram character-level tf-idf. These features are stacked into a vector and referred to as *count vectors* for the remainder of this paper. Secondly, stylistic features use natural language techniques to gain grammatical information to better understand the syntax and style of the document. Lastly, psychological features are closely related to emotions and the cognitive aspect of natural language processing. The psychological features are researched further by utilizing various features including; VAD *(Valence, Arousal, Dominance)* (Warriner et al., 2013), emotional embeddings, and LIWC (Tausczik and Pennebaker, 2010). The features are run in combination with each other in various experiments to test if the top features are sufficient enough to detect sarcasm, or if all the features work best to break down the barriers of sarcasm detection.

5.1 LIWC

LIWC (Tausczik and Pennebaker, 2010) is a text analysis program with a built-in dictionary that counts words in psychologically meaningful categories. After all the words have been reviewed, the module calculates the total percentages of words that are similar and match that of the user dictionary categories. LIWC is used to extract features to detect and categorize the meaning, emotional sentiment, and social relationship of the words in the data set.

5.2 Valence, Arousal, Dominance (VAD)

VAD *(Valence Arousal Dominance)* (Warriner et al., 2013) includes almost 14,000 lemmas rated on a 1-9 scale according to the emotions evoked by the terms. Valence refers to the pleasantness of the word, arousal determines how dull or exciting the emotion is, and dominance ranges from submission to feeling in control. The VAD dimensions allow the research to further explore the effective meanings of tweets and threads and determine their viability as a predictor of sarcasm. The VAD scores for each response were computed and then used the three scores were used a feature in the classifier. The scores were used to measure the congruity between the response and context. The VAD scores were calculated for each individual response and context and then subtracted the response scores by their respective counterpart of context score. For

example, if a response receives a valence score of 9 and its corresponding context/0 receives a valence score of 1, the valence congruity score would be a 8. It is hypothesized that sarcastic tweets/threads may show very little effective congruity compared to their non-sarcastic correspondent.

5.3 VADER

VADER *(Valence Aware Dictionary and sEntiment Reasoner)* (Hutto and Gilbert, 2015) is a lexicon and rule-based library built especially for sentiment analysis of social media platforms and their text. VADER maps lexical features to emotions and thus provides insight into the level of such emotions through a series of polarity scores. VADER considers capitalization, punctuation, degree modifiers, emojis, and negations to compute the negative, positive and neutral scores. VADER's compound score provides a normalized, weighted composite score for a given tweet and thread.

5.4 Emotional Embeddings

label	context0_emotion / response_emotion	anger	fear	joy	love	neutral	sadness	surprise
NOT_SARCASM	anger	221	35	176	11	0	69	6
	fear	48	27	51	1	0	25	4
	joy	279	72	898	34	0	131	20
	love	4	2	48	7	0	5	2
	sadness	86	24	101	5	0	60	9
	surprise	16	3	13	0	0	4	3
SARCASM	anger	498	120	345	17	0	116	26
	fear	83	33	61	3	0	12	6
	joy	333	72	289	13	1	96	22
	love	9	3	6	0	0	3	1
	sadness	98	24	88	2	0	41	3
	surprise	30	10	28	2	0	5	1

Figure 1: Distribution of Emotions for Response vs. Context/0 in the training data for Twitter

The emotions conveyed in the data set are portrayed through emotional embeddings. Calculating the emotions of the text goes a level further than just looking at the word embeddings. Using a pretrained model from Hugging Face (Saravia et al., 2018), the tweets are categorized into six emotions. The emotions include, *joy, anger, fear, surprise, sadness* and *love*. Figure 1 above represents an example of the distribution of the emotions between response and context/0 in the balanced training data set from Twitter. The results support the hypothesis of this paper that sarcasm is typically associated with a type of negative emotion. When the context is labeled as "anger", non-sarcastic tweets tend to respond with joy or a positive emotion, while sarcastic tweets usually respond with anger or a negative emotion. In contrary, when the context is labeled as "joy", non-sarcastic tweets overwhelmingly respond with joy, while sarcastic tweets still largely respond with anger. There are 1,216 instances of the same emotion expressed in both response and context for the non- sarcasm class and 863 instances of this in the sarcasm class. Sarcastic tweets are generally in-congruent with emotions throughout the response and context, unless associated with a negative emotion, e.g., *anger*. The results from Reddit were similar as well. A negative thread that was labeled sarcastic was typically met by a negative, anger, corresponding context. This helps prove the intuition that a sarcasm text in social media, is typically identified as a negative emotion and also matched and responded with by a negative or angry emotion. While in contrast, non-sarcastic were typically more positive and joyful.

5.5 Context Similarity Scores

The standard document similarity estimation technique using word embeddings (GloVe, Pennington et al. 2014) and emotional embeddings was used to calculate the cosine similarity. (Saravia et al. 2018), which consists of measuring the similarity between the vector representations of the two documents. Let $x_1, \ldots x_m$ and $y_1, \ldots, y_n$ be the emotion (or word embedding) vectors of two documents. The cosine similarity value between the two documents (e.g., a tweet and its context) centroids $C_x = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $C_y = \frac{1}{n}\sum_{i=1}^{n} y_i$ is calculated as follows:

$$\cos(C_x, C_y) = \frac{\langle C_x, C_y \rangle}{\|C_x\| \|C_y\|}, \qquad (1)$$

where $\langle x, y \rangle$ denotes the inner product of two vectors x and y.

Two cosine similarity scores are computed: 1) semantic cosine similarity using word embeddings; 2) cosine similarity using emotional embeddings. The linguistic intuition is that a sarcastic response is going to be semantically or emotionally in-congruent with its context. This in-congruence, creates the presence and shows the effect of sarcasm.

	Message
c/0	It's no secret that this president has routinely targeted religious and ethnic minorities. He has fanned the flames of hate against refugees, Muslims, Africans, immigrants, women and all racial and religious minorities.
c/1	He is routinely and openly hostile to any legitimate Congressional oversight. He has made clear his wanton corruption by soliciting a bribe from a foreign government for his personal political gain.
R	Yassss queen, you're so brave and bold.

Table 2: Sarcastic Tweet.
Response=R; Context0=C/0; Context1=C/1

	Message
c/0	A2 I revert back to Canvas. I am sure you can post assignments for parents in this, (haven't done this yet). Canvas = #thebomb #KidsDeserveIt
c/1	Can you telk me more about Canvas? I haven't heard of it.
R	It's Edmodo with #MorePower You can create assignments in it, post all work, the assignments can be auto graded and imported into your Skyward grade book.

Table 3: Non Sarcastic Tweet.
Response=R; Context0=C/0; Context1=C/1

Table 2 is an example of a sarcastic tweet whose context/0, context/1 and response received an emotion of anger, anger, and joy, respectively. Table 3 represents a non-sarcastic thread of tweets where each message was classified as joy. This indicates that non-sarcastic tweets tend to be more emotionally similar to the preceding context while sarcastic tweets tend to shift in emotion. As a result, when compared to its contexts, the sarcastic tweet received lower emotional similarity scores than the non-sarcastic tweet.

5.6 Feature Analysis

After running all of the features on the training data, SHAP *(SHapley Additive exPlanations)* (Lundberg and Lee, 2017) was implemented to determine which features are the most important for classification. SHAP is a theoretic output technique that explains predictions of the model, by producing a SHapley score that plots the most important features in our model. The features produced by SHAP were used in the experiments and are referred to as the "select linguistic features". The top 20 features SHAP selects contain a combination of character features such as character count, as well as a number of sentiment features, including VADER scores, emotion scores for both a response and its context as well as VAD features. SHAP was used solely to test the hypothesis that computational sarcasm detection requires the most linguistic features possible, rather than just the top select features. These top scores were run separately and compared to all features run simultaneously.

6 Experimental Evaluation

6.1 Data Preprocessing

The preprocessing procedure consists of steps to remove noisy and unnecessary data. First, the data is tokenized (Loper and Bird, 2002). Secondly, removed any instance of "@USER" due to the repetition of this token in the beginning of most tweets. Prior research on the shared task demonstrated that the classifiers did not seem to benefit from the additional context. After a thorough investigation of both data sets, it was noticed that a majority of the responses only contained context/0 and context/1. Therefore, any context after context/1 was dropped from both data sets. Also after reviewing the data, it was decided to not remove any stop words, due to the small amount of text in each corpus. Eliminating the stop words would have removed crucial information to the data and leaving too little text to work with. Lastly, all punctuation and emojis were left, as they proved to be useful information during the extraction and usage of certain features. VADER takes into consideration the punctuation and the emojis into its classification.

7 Results

A Random Forest classifier was used on different experiments of which the most relevant ones are outlined in Table 4. The baseline scores represent an attention based LSTM model described in Ghosh et al. (2018) and used in the CodaLab Shared Task. It is viewed how each feature performed on just the response versus the response and context. For Twitter, noticeably, a combination of all count features and all linguistic features achieves the best F1 score of 67%. This score is further increased to 70% when context is considered through the classifier. For Reddit, a combination of all linguistic features received a 56% recall, while increasing to 64% while considering the context. These results confirm the hypothesis that all linguistic features as well as the context, to achieve the highest results. Disregarding the contextual information in tweets and threads loses information that is crucial for computational automatic sarcasm detection. One last observation is that, Twitter

achieved a higher accuracy score in comparison to Twitter, indicating that presence of hashtags aides in sarcasm detection with Twitter having over 2,000 more hashtags than Reddit.

Experiments	P	R	A	F1
Baseline 1 Shared Task	70%	**66.9%**	N/A	68%
T:R(Sel. Ling Ft.)	54%	60%	59%	52%
T:R(All Ling Ft.)	70%	64%	65%	67%
T:R+C/0+C/1(Sel. Ling Ft.)	71%	60%	62%	64%
T:R+C/0+C/1(All Ling Ft)	**80%**	62%	**66%**	**70%**
RD:R(Sel. Ling. Ft.)	59%	54%	60%	60%
RD:R(All Ling. Ft.)	59%	56%	60%	52%
RD:R+C/0+C/1 (Sel. Ling. Ft.)	52%	45%	53%	56%
RD:R+C/0+C/1 (All Ling. Ft.)	60%	64%	60%	53%

Table 4: Random Forest; Various feature combinations. *Response=R; Context0=C/0; Context1=C/1; Count=Ct; Linguistics= Ling; T:=Twitter experiment RD:= Reddit experiment Features= ft.*

8 Conclusion

In this paper, the role various linguistic features play in computational sarcasm detection are explored. Further, it is investigated by combination of text and word count features, stylistic and psychological features. The result of the experiments indicate that contextual information is crucial for sarcasm detection. It is also observed that sarcastic tweets are often incongruent with their context in terms of sentiment or emotional load. Using a Random Forest classifier and the features that were extracted, obtain promising results. The presence of hashtags appears to be a imperative feature in sarcasm detection as well. Current work includes the exploration of the linguistic features utilized in this paper, in combination with a transformer-based architecture. Lastly, to test experiments on different social media platforms to see if the presence of all features and the presence of context is crucial.

References

Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. 2020. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87.

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland. Association for Computational Linguistics.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491.

Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012, Lisbon, Portugal. Association for Computational Linguistics.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, Oregon, USA. Association for Computational Linguistics.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2015. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 25–30.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rishabh Misra and Prahal Arora. 2019. Sarcasm detection using hybrid neural network. *ArXiv*, abs/1908.07414.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 97–106.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*. The AAAI Press.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 765–770. IOS Press.

Byron C. Wallace, Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China. Association for Computational Linguistics.

Amy Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45.