Linguistic Features in Automatic Sarcasm Detection
A Comparison 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. This paper is an expansion up on the work done solely on Twitter by Ducret et al. (2020). The main goal is to divulge deeper upon the experiments previously done by Ducret et al. (2020) and test the classifiers results on Reddit while also incorporating the presence of other character counts, such as hashtags. This paper also examines the lexical diversity of both platforms. The linguistic features that are investigated are a combination of count, stylistic, and psychological features. The features are first run on a corpus of tweets and then separately on a corpus of 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. Secondly, sarcastic responses are typically associated with a negative emotion; 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. Lastly, a lower lexical diversity also seems to aide in the experiments of sarcasm detection.

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.

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.

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, machine learning models and transformer-based-architectures have been used to achieve significant results. 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.

Like most sentiments and emotions, sarcasm is neither positive or negative, which make sarcasm difficult to define. The goal is to investigate which linguistic features play the largest role. Furthermore, investigate if only utilizing the top features are enough to detect sarcasm. Due to the fact the borders 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 both social media platforms are compared and contrasted with each other. While this paper only researches two social media accounts, 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

Automatic sarcasm detection is a highly active field, especially in the fields on NLP and computational work. 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. As mentioned previously, this research is an expansion upon the work done on the Twitter corpus and the experiments run by Ducret et al. (2020). Their work indicates promising results utilizing the features and run through the classifier. The work of Ducret et al. (2020) indicates that contextual information is completely necessary in sarcasm detection. This work is to test the results on a different social media platform, with similar contextual information. Both corpora are similar in regards to contextual information and lexical diversity, which is discussed later in this paper.

3 Approach

The approach this paper takes utilizes a combination of count, stylistic, and psychological linguistic features to automatically detect the presence or absence of sarcasm in a given text. This approach intentionally experiments with the same classical machine learning algorithm used in Ducret et al. (2020) 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. To avoid any noisy data, retweets, duplicates of tweets and threads, and quotes have been deleted and excluded from the data set used. Please visit Ghosh et al. 2020 for more information on the shared task and data set as well as Ducret et al. (2020) for the original project performed merely on Twitter. 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 applicable). The research done in this paper only explores the response, context/0 and context/1 for both accounts.

4.1 Lexical Diversity

Type token ratio was calculated to investigate each corpus separately. Type token ratio is computed to investigate the vocabulary variation in a given text. The types refer to the total amount of *different* words, while the tokens refer to the *total* amount of words in the corpus. The type token ratio is calculated by dividing the types occurring in the corpora by its tokens. As seen below, Twitters corpus includes 14,406 types and 153,047 tokens. This brings the TTR for Twitter to 9.4%. Reddit having 13,234 types and 101,461 tokens, bringing the TTR rate of Reddit to 13%. Both of these social media platforms indicate a very low lexical diversity. A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite, a low level of vocabulary variation. The educated speculation is that the lower the TTR, the better the corpora will perform in sarcasm detection. Certain texts such as scholarly articles, will have a much higher type token ratios, most likely indicating to perform poorly in sarcasm detection. The count of each hashtags were also calculated on each forum separately to investigate any similarities and differences. Twitters corpus shows an overwhelming amount more of hashtags in its data set than Reddit. The linguistic hypothesis is that the presence of more hashtags will perform better through the classifier.

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 *count vector*. Ducret et al. (2020). Secondly, stylistic features use natural language processing techniques to obtain grammatical, semantic, and syntactical information to better understand the methodology 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. Previous work done on Twitter, shows promising results that the use of all additional linguistic features are necessary and aide in sarcasm detection Ducret et al. (2020).

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.

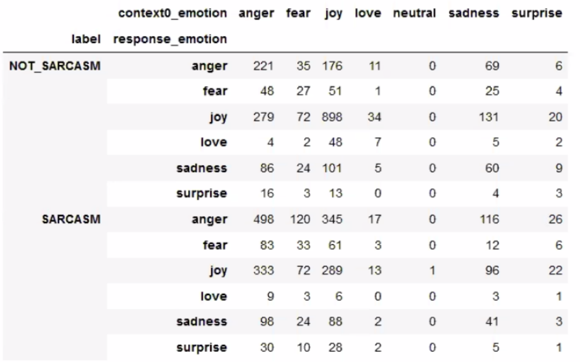
5.2 Valence, Arousal, Dominance (VAD)

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



	context0_emotion	anger	fear	joy	love	neutral	sadness	surprise
label	response_emotion							
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 from Ducret et al. (2020)

The emotions conveyed in the data set are portrayed through emotional embeddings. Calculating the emotions of the text explores a level further than just obtaining the word embeddings of a given text. Using a pre-trained model from Hugging Face (Saravia et al., 2018), the tweets and threads are categorized into six emotions. Those six 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 Ducret et al. (2020). The results from previous work completed on Twitter Ducret et al. (2020), indicate 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 display similar outcomes. A negative thread that was labeled sarcastic was typically met by a negative corresponding context. This evidence from a second social media account 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

A similarity algorithm 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.

The same two cosine similarity scores were computed from the experiment done on Twitter Ducret et al. (2020): 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 and or emotionally in-congruent with its contextual counter part. These findings need to be proven true after being run on Reddit, to hold validity.

5.6 Feature Analysis

After performing all of the features on the training data on both Twitter and Reddit separately, SHAP *(SHapley Additive exPlanations)* (Lundberg and Lee, 2017) was implemented to determine which features are the most important for classification. SHAP provides an output that explains the predictions of the model. This technique is used to perform a feature analysis on the data, and then utilize the top twenty features from SHAP, and implement those select features on the testing data. The features produced by SHAP are referenced in this paper as experiments run on "select linguistic features". After reviewing the features selected by SHAP, the features were a combination of count features, psychological features (including emotional embeddings and LIWC) as well as stylistic features. Results from Ducret et al. (2020) express that these select linguistic features will not out perform the utilization of all the features. SHAP was used solely to test the hypothesis and the results of the previous work done. 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 set of characters in the beginning of most tweets. Prior research completed on the shared task demonstrated that the classifiers performed by others 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. Additionally 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

The features were run through various experiments and combinations in a random forest classifier. The most relevant combinations are referenced in Table 2. 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% when only taking into consideration the response. This score is increased to 70% when context is included through the classifier Ducret et al. (2020). For Reddit, a combination of all linguistic features received a 56% recall, while increasing to 64% when considering the context. These results confirm the hypothesis that all linguistic features as well as the context are necessary to achieve the highest results. Disregarding the contextual information in tweets and threads loses facts and certain knowledge that is crucial for computational automatic sarcasm detection. Additionally, Twitter achieved a higher accuracy score in comparison to Reddit, indicating that presence of hashtags aides in sarcasm detection. Twitter having over 2,000 more hashtags than Reddit proves this theory to be true. The last hypothesis is also confirmed by the results in that Twitter performs slightly better than Reddit, having a slightly lower lexical diversity by 3.6%. A corpus of higher lexical diversity may perform sub-standardly in comparison.

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 2: 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 prove and gain more insight of the prior work done on solely Twitter Ducret et al. (2020). Contextual information is crucial for sarcasm detection. It is also observed that sarcastic tweets are often incongruent with their context in terms of sentiment and their emotions. Using a Random Forest classifier and the features that were extracted, obtain promising results on both forums. 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. Additionally, to test experiments on different social media platforms to see if the presence of all features and the presence of context is crucial. Lastly, further research would be conducted to gain more insight of where and when the data sets were extracted from. Investigating the subject of tweets/threads and the timeline of when they were released may change the context of the data. Advanced research in this field would be to test on Twitter and Reddit again, while having a larger data set with more diverse contextual information.

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.

Martina Ducret, Lauren Kruse, Carlos Martinez, Anna Feldman, and Jing Peng. 2020. You don't say... linguistic features in sarcasm detection. *CEUR Workshop Proceedings*, 2769. Publisher Copyright: Copyright © 2020 for this paper by its authors. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.; null ; Conference date: 01-03-2021 Through 03-03-2021.

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

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.

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

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.

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.

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.

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.

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