What A Sunny Day ☔: Toward Emoji-Sensitive Irony Detection

Aditi Chaudhary* **Shirley Anugrah Hayati*** **Naoki Otani*** **Alan W Black**
Language Technologies Institute
Carnegie Mellon University
{aschaudh, shayati, notani, awb}@cs.cmu.edu

Abstract

Irony detection is an important task with applications in identification of online abuse and harassment. With the ubiquitous use of non-verbal cues such as emojis in social media, in this work we study the role of these structures in irony detection. Since the existing irony detection datasets have <10% ironic tweets with emoji, classifiers trained on them are insensitive to emojis. We propose an automated pipeline for creating a more balanced dataset.

1 Introduction

Social media text often contains non-verbal cues, such as emojis, for users to convey their intention. Statistics have shown that more than 45% of internet users in the United States have used an emoji in social media[1]. Due to this prevalent usage of emoji, some works attempt to exploit the occurences of emoji for tackling NLP tasks, such as sentiment analysis (Chen et al., 2019), emotion detection, and sarcasm detection (Felbo et al., 2017), as the presence of emoji can change the meaning a text as an emoji can have positive or negative tone.

We are interested in analyzing the role of emoji in irony since this specific linguistic phenomenon is related to sentiment analysis and opinion mining (Pang et al., 2008). Irony can also relate to more serious issues, such as criticism (Hee et al., 2018) or online harassment (Van Hee et al., 2018). Based on our analysis on existing irony dataset from SemEval 2018 (Van Hee et al., 2018), only 9.2% of the ironic tweets contain an emoji. Furthermore, they crawled tweets using irony-related hashtags (i.e. #irony, #sarcasm, #not). This does not capture all variations of ironic occurrences, especially those caused by emojis.

How an emoji changes the meaning of irony tweets is illustrated by the following example. If we have this tweet: "What a sunny day ☀", it does not sound ironic. However, "What a sunny day ☔" is ironic. From these examples, we can see that we cannot ignore the importance of emoji to identify irony.

Due to the sparcity of ironic tweets containing emoji, our goal is to augment the existing dataset such that the model requires both textual and emoji cues for irony detection.

We first analyze the behavior of emojis in ironic and non-ironic expressions. We find that the presence of emojis can convert a non-ironic text to an ironic text by causing sentiment polarity contrasts. We develop heuristics for data augmentation and evaluate the results. Then, we propose a simple method for generating ironic/non-ironic texts using sentiment polarities and emojis.

2 Related Work

A common definition of verbal irony is saying things opposite to what is meant (McQuarrie and Mick, 1996; Curcó, 2007). Many studies have diverse opinions regarding sarcasm and irony being different phenomenon (Sperber and Wilson, 1981; Grice, 1978, 1975) or being the same (Reyes et al., 2013; Attardo et al., 2003). In this work, we do not make a distinction between sarcasm and irony.

Previous work on irony detection relied on hand-crafted features such as punctuation and smiles (Veale and Hao, 2010) or lexical features, such as gap between rare and common words, intensity of adverbs and adjectives, sentiments, and sentence structure (Barbieri and Saggion, 2014).

More recently, Van Hee et al. (2016) explore constructions of verbal irony in social media texts, reporting that detection of contrasting polarities is a strong indicator and use sentiment analysis

* equal contributions

[1] https://expandedramblings.com/index.php/interesting-emoji-statistics/

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 212–216

Train	All	Ironic				Non-Ironic
		Irony	1	2	3	
# Tweets	3817	1901	1383 (73%)	316 (17%)	202 (10%)	1916
# Tweets containing emoji	406	175	162 (93%)	7 (4%)	6 (3%)	231
# Unique emojis	158	104				122
Test						
# Tweets	784	311	164 (53%)	85 (27%)	62 (20%)	473
# Tweets containing emoji	88	33	27 (82%)	3 (9%)	3 (9%)	55
# Unique emojis	81	23				70

Table 1: Dataset statistics

Ironic		Non-ironic	
Emoji	**Count**	**Emoji**	**Count**
😂	42 (29.0%)	😂	49 (31.2%)
😒	26 (17.9%)	🤷‍♀️	22 (14.0%)
😳	12 (8.3%)	😘	16 (10.2%)
🙈	11 (7.6%)	😁	14 (8.9%)
😑	10 (6.9%)	😡	14 (8.9%)
😍	9 (6.2%)	💜	11 (7.0%)
😭	9 (6.2%)	👌	11 (7.0%)
🔫	9 (6.2%)	❤️	11 (7.0%)
👊	9 (6.2%)	👊	9 (5.5%)
👍	8 (5.5%)	👍	8 (4.8%)

Table 2: Top 10 most frequent emojis in ironic tweets and non-ironic tweets along with the count and percentage of each emoji.

for the same. Machine learning algorithms such as SVMs informed with sentiment features have shown good performance gains in irony detection (Van Hee, 2017, 2018).

Some neural network-based methods have been conducted. LSTM has proven to be successful for predicting irony. Wu et al. (2018), that ranks first for SemEval 2018: Shared Task on Irony in English Tweets Task A, utilizes multitask-learning dense LSTM network. The second-ranked participants, Baziotis et al. (2017), uses bidirectional LSTM (biLSTM) and self-attention mechanism layer. Ilić et al. (2018)'s architecture is based on Embeddings from Language Model (ELMo) (Peters et al., 2018) and passes the contextualized embeddings to a biLSTM. Ilić et al. (2018)'s model becomes the state of the art for sarcasm and irony detection in 6 out of 7 datasets from 3 different data sources (Twitter, dialog, Reddit).

3 Proposed Approach

3.1 Dataset Analysis

We analyze the SemEval 2018: Irony Detection in English Tweets dataset (Van Hee et al., 2018). Table 1 shows the data statistics for both ironic and non-ironic tweets. Row 2 shows the tweet distribution with respect to the presence of emojis. We can see that only 11% of the all tweets contain an emoji, out of which 46% are ironic. In order to study the robustness of current irony detection model to ironic text containing emoji, it is necessary to augment the existing dataset with additional tweets containing emojis due to the limited amount of ironic tweets with emojis.

We hypothesize that the emojis used for ironic tweets may be different from the emojis used for non-ironic tweets. Table 2 shows ten emojis that most frequently appear in ironic tweets and non-ironic tweets in the English dataset. 😂 appears most often in both ironic (42 times) tweets and non-ironic (49 times) tweets. For other frequent emojis, except for 👍, the emojis in ironic tweets are different from the emojis in non-ironic tweets. Emojis in the ironic tweets mostly does not have positive sentiment, if we do not want to say negative, such as 😑, 😭, 😒, 😩, and 🔫 while the most frequent emojis in the non-ironic tweets are dominated by positive emojis, such as 😘, 😁, 💜, 👌, and ❤️. Moreover, some tweets may contain multiple emojis. We found that out of 175 ironic tweets that contain emoji, 45% of them contain multiple emojis. We consider to follow this distribution when we are building our ironic tweet generation pipeline.

3.2 Manual Data Augmentation

To further analyze the role of emoji in ironic expressions, we conduct qualitative analysis while controlling the effect of the text content. Concretely, we generate ironic and non-ironic texts by manipulating emoji without changing the texts. The resulting texts give us an insight about emoji use and can also be used as an evaluation resource for developing emoji-sensitive irony detection.

Our manual inspection focuses on the three cases of emoji manupulation below.

1. **Case 1 - Irony with emoji → non-irony:** We randomly sample 50 ironic tweets containing

	Original Example	Transformed Example
Case 1	My year is ending perfectly 😂 (Ironic)	My year is ending perfectly 😊 ❤️ (Non-Ironic)
Case 2	Finally went to the doctor and feeling so much better. (Non-Ironic)	Finally went to the doctor and feeling so much better 👎🏽 . (Ironic)
Case 3	Another day in paradise haha (Ironic)	Another day in paradise haha 😘 ❤️ (Non-Ironic)

Table 3: Examples of annotated tweets with respect to the different cases.

emojis from the original dataset and inspect whether replacing/removing the emojis converts these ironic tweets to non-ironic tweets.

2. **Case 2 - Non-irony without emoji → irony:** We randomly sample 50 non-ironic tweets without containing emojis and inspect whether adding emoji turns these non-ironic tweets to ironic tweets.

3. **Case 3 - Irony without emoji → non-irony:** For another set of randomly sampled 50 ironic tweets not containing any emojis originally, we inspect whether addition of any emojis converts these ironic tweets to non-ironic tweets.

For each original tweet in each case, three annotators assign a label '1' in case a conversion is possible and '0' otherwise. Additionally, for tweets that can be converted, each of the annotators provides one transformed tweet. Table 3 shows some example tweets. After this annotation step, we obtained 171 transformed tweets in total.

We calculate the inter-annotator agreement for each case in Table 4. Case 2 has the worst agreement. This is possibly because it is difficult to convert non-ironic tweets to ironic tweets only by adding emoji.

For instance, two out of the three annotators felt the following non-ironic tweet "*@MiriamMockbill must b in the #blood lol x*" can be transformed into an ironic tweet "*@MiriamMockbill must b in the #blood lol x* 😆" by adding emojis, however the irony in the transformed tweet is not very evident.

Next, we validate the quality of the generated texts. Each example of the generated texts is given to the two annotators. The annotators must rate the given example as 'ironic' or 'non-ironic'. The agreement was moderately high. We achieved 100% agreement on 100 out of 171 tweets (58.4%). We call this dataset consisting of the generated 100 tweets plus their 60 original tweets **Imoji** dataset and use it in a subsequent analysis. To the best of our knowledge, this is the first dataset which contain multiple ironic/non-ironic expressions with the same text body and different emojis.

	Fleiss' κ	% Agreement
Case 1	0.49	62%
Case 2	0.02	30%
Case 3	0.23	52%

Table 4: Fleiss' κ and percent agreement scores for calculating inter-annotator agreement.

3.3 Automatic Data Augmentation

Analysis of Imoji dataset suggests that emojis tend to be used for causing "irony by clash" in most cases. Positive emoji is likely to be paired with negative texts in ironic expressions, and vice versa.

3.3.1 Method

Following the insight drawn from Imoji dataset, we propose a simple data augmentation using sentiment analysis dataset so that we can build an ironic detector more robust to emoji.

1. We collected emoji-sentiment lexicon from Emoji Sentiment Ranking (Kralj Novak et al., 2015). This resource contains the emojis' frequency and sentiment polarity. Then, we preprocessed the emojis in from this Emoji Sentiment Ranking, resulting in 48 strongly positive 48 emojis and 48 strongly negative emojis. We filter out low-frequency emoji (bottom 50% frequency), ignore non-emotional symbols (e.g. arrows), and extract top 10% emoji in terms of normalized sentiment scores for each of positive and negative sentiments.

2. Collect tweets with positive and negative sentiments from SemEval 2018 Affect in Tweets Task 3 dataset (Mohammad et al., 2018). This dataset contains total of 2,600 tweets with negative emotions, such as sadness, anger and fear, joy tweets, and sarcastic tweets. Crowdsourcers were asked to annotate them as positive or negative tweets.

3. Generate ironic/non-ironic tweets by adding emoji at the end of texts.

Text	Tweet	Emoji	Label
now that I have my future planned out, I feel so much happier #goals #life #igotthis #yay 🌟	+	-	Ironic (Yes)
Never let me see you frown ❤️	-	+	Ironic (?)
MC: what are you listen to these days?Bogum: these days I feel gloomy, I listen to ccm (spiritual song) often. Church oppa mode. :) 😍	-	+	Ironic (No)
Love your new show @driverminnie 💝	+	+	Non-ironic (Yes)

Table 5: Generated ironic examples. Tweet refers to tweet sentiment and emoji refers to emoji sentiment

3.3.2 Evaluation of Automatic Generation

We conduct manual analysis of the generated tweets. Table 5 displays the generated ironic and non-ironic tweets.

The first example is generated by combining positive sentiment tweet with negative sentiment emoji, and we agree that it is an ironic text. For the second example, it is quite unclear whether the text is ironic or not. ❤️ may not make the text ironic if the writer's purpose is really not to see the other person frown even though the sentiment of the text without emoji itself is slightly negative. The third example is not ironic although it is generated by combining negative tweet with positive emoji. "Bogum" is a Korean actor and "oppa" is commonly used by fangirls to call older Korean male. Thus, using 💝 in the text makes sense and does not make it ironic. The last example is a generated non-ironic text by adding positive emoji to positive tweet. Based on this analysis, we decided to use only tweets with positive sentiments as seeds to generate accurate ironic/non-ironic tweets.

4 Experiments

4.1 Preprocessing

To normalize special strings in tweets like URLs, mentions and hashtags, we run ekphrasis[2] (Baziotis et al., 2017) to normalize texts. We also correct non-standard spellings. We collect sentiment analysis datasets for automatic data augmentation from SemEval 2018 Shared Task (Mohammad et al., 2018). Then we obtained 768 additional irony detection instances.

4.2 Baseline Model

We use the NTUA-SLP system (Baziotis et al., 2018) from SemEval 2018. It uses standard two-layer biLSTMs and a self-attention mechanism to encode a tweet into a fixed-sized vector and makes a prediction by a logistic regression classifier taking the encoded tweet as input. Embedding layers are initialized with 300D pre-trained word embeddings, word2vec model trained on tweets for English ((Baziotis et al., 2017)).

[2]https://github.com/cbaziotis/ekphrasis

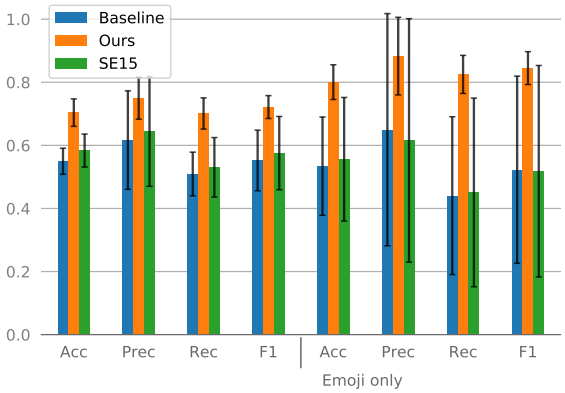

Figure 1: Result on irony detection in Imoji dataset. Baseline referes to SemEval 2018 train set, ours is baseline and our generated data (+767 instances), SE15 is baseline and SemEval 2015 dataset (+767 instances). Performances are mean averages over 10 trials, and error bars denote standard deviations.

4.3 Result

We train the model on our augmented data and test it on the Imoji dataset as shown in Figure 1. To make sure that the performance change by our augmented data (Ours) is not only from the increased number of training instances, we also collect the same number of ironic detection instances as the generated instances from another dataset containing irony annotations (Ghosh et al., 2015). Interestingly, the classifier trained on our augmented dataset achieve much higher recall.

5 Conclusion

In this work, we presented an automatic pipeline for generating ironic data using sentiment analysis. We observe that our method works well for the irony based on polarity contrast. In summary, the experimental results show our augmented data helped classifiers improve their sensitivity to emojis in irony detection tasks without damaging the overall performance of irony detection on the whole datasets. An interesting future direction is to apply our method to multilingual irony dataset.

References

Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260.

Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in twitter. In *Proceedings of the Student Research Workshop at EACL*, pages 56–64.

Christos Baziotis, Athanasiou Nikolaos, Pinelopi Papalampidi, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns. In *Proceedings of SemEval*, pages 613–621.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of SemEval*, pages 747–754.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-lingual sentiment classification. In *The World Wide Web Conference*, pages 251–262. ACM.

Carmen Curcó. 2007. Irony: Negation, Echo, and Metarepresentation. *Irony in Language and Thought*, pages 269–296.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of SemEval*, pages 470–478.

H Paul Grice. 1975. Logic and conversation. speech acts, ed. by peter cole and jerry morgan, 41-58.

H Paul Grice. 1978. Further notes on logic and conversation. *1978*, 1:13–128.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Exploring the fine-grained analysis and automatic detection of irony on twitter. *Language Resources and Evaluation*, 52(3):707–731.

Suzana Ilić, Edison Marrese-Taylor, Jorge A Balazs, and Yutaka Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. *arXiv preprint arXiv:1809.09795*.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS ONE*, 10(12):e0144296.

Edward F McQuarrie and David Glen Mick. 1996. Figures of rhetoric in advertising language. *Journal of Consumer Research*, 22(4):424–438.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of SemEval*, pages 1–17.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Cynthia Van Hee. 2017. Can machines sense irony?

Cynthia Van Hee. 2018. Exploring the fine-grained analysis and automatic detection of irony on twitter. In *Proceedings of LREC*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Exploring the realization of irony in twitter data. In *Proceedings of LREC*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of SemEval*, pages 39–50.

Tony Veale and Yanfen Hao. 2010. Detecting Ironic Intent in Creative Comparisons. In *Proceedings of ECAI*, pages 765–770.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In *Proceedings of SemEval*, pages 51–56.