Unmasking Sarcasm - Enhancing Sentiment Analysis in E-Commerce Reviews and Questions

Hugo Bouy
Illinois Tech
hbouy@hawk.iit.edu

Rémi Kalbe
Illinois Tech
rkalbe@hawk.iit.edu

Mathias Roumane
Illinois Tech
mroumane@hawk.iit.edu

1. Problem description

In an increasingly e-commerce driven world, the analysis of review, comments and questions written by consumers online has become a field of interests. Understanding opinions and emotions expressed in product feedback/related questions is essential to enhance the user's experience. One aspect that might be overlooked in this area is sarcastic and humor detection that can lead to a misinterpretation of these texts.

Humor remains a complex human phenomenon that is far from having a clear definition. While humor and sarcastic mechanisms are integral to human interaction, its subjective nature makes it a challenging target for computational analysis. Recent work has been able to open up this area using deep learning and natural language processing advances. With this project, we will attempt to improve e-commerce review processing using deep learning models for humor disambiguation.

2. Brief Survey of Previous Work

Several studies have been conducted over the past years in the ambition of detecting humor and sarcasm:

Jain et al. [5] delved into the complexities of identifying sarcasm in Amazon reviews. Recognizing sarcasm is crucial for accurate sentiment analysis, especially since sarcastic comments can be misinterpreted by traditional opinion mining methods. They utilized the "Sarcasm Corpus" containing labeled ironic and regular Amazon reviews, extracting features like sentiment scores, punctuation patterns, and contextual elements that consider the contrast between review sentiment and product rating. Their experiments designated the Support Vector Machine (SVM) classifier as the most accurate, emphasizing the role of context in sarcasm detection.

Building upon the idea of sarcasm detection, Poria et al. [8] introduced a method using deep convolutional neural networks (CNNs). They critiqued traditional methods that treat sarcasm detection as mere text categorization, arguing that such approaches often miss the deeper understanding of language nuances required for sarcasm. Their method integrates sentiment, emotion, and personality features extracted from pre-trained CNNs. By leveraging Twitter data, they contrasted sarcastic sentences with the ground-truth polarity of events. Their experiments with word embeddings from word2vec and a combined CNN-SVM approach demonstrated superior performance on benchmark datasets.

Yaghoobian et al. [9] further discussed the challenges of sarcasm detection in sentiment analysis. They categorized detection methods into content-based, which focus on lexical indicators, and context-based, which emphasize background knowledge. Their study highlighted the CASCADE model, which uses user embeddings to capture user-specific features, as an example of leveraging context for sarcasm detection.

Shifting the focus to humor detection, Ziser et al. [10] identified product bias in Product Question Answering (PQA) systems, where certain products attract more humorous questions. They proposed a deep-learning framework to detect humor in PQA, focusing on incongruity and subjectivity.

Annamoradnejad and Zoghi [1] proposed the ColBERT model for humor detection. This model leverages BERT embeddings for sentence representation and has achieved state-of-the-art results on various datasets.

Lastly, Gupta et al. [4] explored the potential of Large Language Models (LLMs) in humor detection. Their research emphasized the capability of LLMs to capture the intricacies associated with humor and offense detection.

3. Datasets

3.1. Combined dataset for sarcasm detection

To train our future models, we built a collection of datasets containing several sarcastic and non-sarcastic text. The main datasets used in the scope of this project are listed below.

- **Headlines dataset [6]**: Contains a list of 28,619 headlines collected from two news websites. On one hand,

TheOnions aims to produce sarcastic versions of real news events. On the other hand, real and non-sarcastic news headlines are collected from HuffPost. This dataset has the advantage of having no spelling mistakes and informal usage since it is written by dedicated professionals in a formal manner.

- **MUStARD++ dataset [2]**: Mustard++ is a multimodal sarcasm dataset that has been annotated with 9 emotions. It was compiled from popular TV shows such as Friends, The Golden Girls or The Big Bang Theory. We will be using this dataset mostly to detect sarcasm but if we have time we may use the annotation to classify the emotion associated with the sarcastic sentence considered.

- **Sarcasm Corpus V2 dataset [7]**: This dataset contains both sarcastic and non-sarcastic utterances. They are additionally classified in three different types: generic (6,520 samples), hyperbole (1.164 samples) and rhetorical (1,702 samples).

- **Sarcasm Amazon reviews dataset [3]**: Contains a large number of both regular and ironic Amazon reviews. Each review is also associated with information about the product for which the review was written, the number of stars assigned by the author, etc. This dataset will be the most useful in the second part of our project when we link the sarcastic review with the information regarding the product.

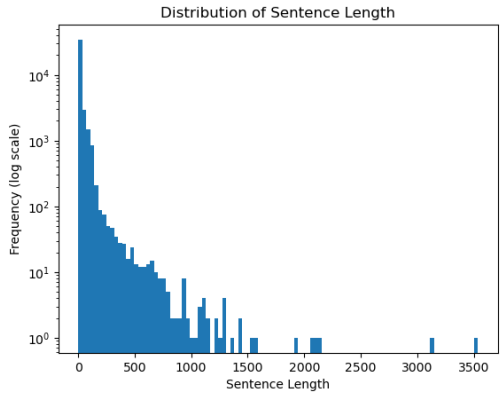

Figure 1. Frequency of sentence length in the final combined dataset

As an initial data processing step, we first concatenated those datasets together in order to use them to implement an initial sarcasm detection model. The final combined dataset then has the full sentence or review and a sarcastic indicator (0 or 1). The final combined dataset contains 40,461 samples from different sources and types. We then have a quite complex dataset with sentences of very variable length as displayed in figure 1.

3.2. Amazon reviews dataset

In order to explore more in depth our problematic around e-commerce, the sarcasm Amazon reviews dataset remains the most useful one in the scope of our project. As described previously, it contains reviews of products on Amazon that are labelled either as sarcastic or regular. Additionally, this dataset reports other information associated to the review such as the product associated with it, the number of stars assigned by the author, etc. One of the sampled sarcastic reviews extracted from this dataset is given in table 1 as an example.

Feature	Example
STARS	3.0
TITLE	Great Product, Poor Packaging
DATE	May 14, 2009
AUTHOR	Patrick J. McGovern ”Procrastinating Evil Scientist”
PRODUCT	Uranium Ore
REVIEW	I purchased this product 4.47 Billion Years ago and when I opened it today, it was half empty.

Table 1. Example of an ironic review sample and the associated information

The additional information offered by this dataset will allow us to take additional parameters into account rather than basing our prediction solely on the review. We consider that adding these features may help the model to better identify the underlying patterns of sarcasm.

4. Models

4.1. Simple Sarcasm detection model

As introduced above, we build a heterogeneous dataset with sarcasm examples from very different sources and context. Our goal here is to experiment building a model that could learn the inner structure of sarcasm from a large amount of data.

Before doing so, and as a starting point for this project, we first aimed at creating a simple model capable of detecting if a sentence contains some from of sarcasm using the Amazon review dataset only. At this stage, the model only takes into account the sentence and no additional factors. We will later add additional relevant factors in detecting sarcastic or humorous patterns such as information on the product considered.

Jain et al. previously conducted an experiment on this dataset [5]. Their idea was to build a feature vector for each Amazon review using various features such as the positive/negative sentiment score, punctuation, part of speech

and bigram analysis. Their results shown very good accuracy with Naïve Bayes (77.50%), a multi-layer perceptron classifier (81%), and SVM (81.5%), with similar precision, recall and f1. To verify the necessity of this feature selection, we created 2 models, one SVM and one LSTM and train them on plain data, without feature selection. SVM used count vectorization and the LSTM used word embedding. Our results showed that SVM performed well with 82% accuracy on the Amazon review dataset. However, digging into the results revealed that the model was more likely to misclassify a sarcasm text with both precision and recall being much under the ones for Regular text. One explanation for this difference may be the small size of the dataset which contains only 1254 texts, with only 437 being sarcastic.

	Precision	Recall	F1-score
Regular	0.84	0.90	0.87
Sarcasm	0.72	0.61	0.66

Table 2. Precision, Recall and F1-score for SVM on the Amazon review dataset with only count vectorization.

The LSTM model on its side less performed with a test accuracy of 67%. A Similar behavior can be observed regarding the sarcastic text for which the model fails to understand the sarcasm pattern in all the cases.

	Precision	Recall	F1-score
Regular	0.81	0.71	0.76
Sarcasm	0.43	0.57	0.49

Table 3. Precision, Recall and F1-score for LSTM on the Amazon review dataset with only count vectorization.

These results confirmed the necessity of feature selection before training the model.

Before applying feature selection we experimented with the state-of-the art model BERT over our entire dataset to observe if the wilder and more diverse example set of sarcasm could be a strong benefit.

4.2. BERT-based Sarcasm Detection Model

4.2.1 Overview of BERT

Bidirectional Encoder Representations from Transformers (BERT) was a significant advancement in the field of natural language processing (NLP), as it leverages the Transformer, a state-of-the-art neural network architecture specifically built to effectively process sequential input. In contrast to conventional models that operate on text unidirectionally (either left-to-right or right-to-left), BERT stands out due to its capacity to comprehend the contextual meaning of a word within a sentence by considering the preceding and succeeding words, thus enabling bidirectional processing of the entire word sequence. The aforementioned feature is especially beneficial in tasks such as sarcasm identification, where the contextual interpretation of words plays a crucial role in determining the presence of sarcasm in a statement.

4.2.2 Model Implementation

In our sarcasm detection task, we utilized a pre-trained BERT model, which we then fine-tuned on a dataset consisting of texts labeled as sarcastic or non-sarcastic. By fine-tuning BERT on our specific dataset, the model could apply its deep, bidirectional understanding of language to effectively distinguish between sarcastic and genuine statements.

4.2.3 Experimental Results

The efficiency of our fine-tuned BERT model in sarcasm detection was significant. The validation loss obtained was 0.386, while the accuracy reached was 87.823% on the test set sampled from our 40,461 text dataset introduced previously. The findings of this study emphasize the proficiency of BERT in effectively categorizing sentences as either sarcastic or non-sarcastic, hence showcasing its sophisticated comprehension of language and contextual nuances.

4.2.4 Conclusions

The efficacy of our BERT-derived model in detecting sarcasm illustrates the robustness of bidirectional contextual analysis in comprehending intricate linguistic patterns. The experimental results demonstrate that BERT is effective in handling delicate natural language processing (NLP) tasks, specifically sarcasm detection, as seen by its high accuracy and minimal validation loss. The aforementioned efficacy highlights the wider possibilities of employing advanced, contextually-aware models in complex language comprehension tasks.

4.3. Sarcasm detection with additional factors

4.4. Comparison with Chat-GPT

In order to enhance the depth of our analysis, we conducted a comparative evaluation between our sarcasm detection model and Chat-GPT, primarily focusing on the GPT-3.5-turbo variant developed by OpenAI. The objective of this comparison was to assess the efficacy of our model in comparison to a widely recognized general-purpose language model.

4.4.1 Methodology

Our approach involved querying Chat-GPT with a set of sentences from our dataset, asking it to determine if each

was sarcastic. The queries were formulated in a simple way, while we recognize that the design of the prompt was not extensively optimized and may need more modification to improve accuracy. This observation suggests the possibility of making enhancements in further experimental endeavors.

4.4.2 Observations & Thoughts

The preliminary results of Chat-GPT on our sarcasm detection test were unsatisfactory, exhibiting an accuracy rate of merely 59.75%. The actual application of Chat-GPT was hindered by the time-consuming nature of its interaction, mostly caused by the response latency and rate limiting of the API. This limitation had a notable impact on the usability of Chat-GPT in many circumstances. Furthermore, the responses generated by Chat-GPT occasionally deviated from the intended aim, most likely related to its broad training and reliance on the provided prompts.

Moreover, the financial aspect associated with the utilization of GPT-3.5-turbo proved to be significant, especially considering its rather limited precision as observed in our conducted evaluations. Taking into account these factors, including the processing requirements and concerns regarding response time, raises doubts about the viability of utilizing this approach for sarcasm detection in e-commerce environments on a big scale or in real-time scenarios. Subsequent examinations may be conducted to assess the potentialities of GPT-4; nevertheless, the substantial cost associated with its utilization constitutes a pivotal aspect requiring careful deliberation.

5. What remains to be done

1. **Feature Integration for BERT Model:** The existing BERT-based model will be improved by integrating supplementary product-specific attributes from the Amazon dataset, including product ratings, names, and other pertinent metadata.

2. **Comparison between BERT and simpler model using feature selection:** Compare the efficiency of the previously introduced BERT model and an SVM model using features derived from the text as suggested by Jain et al. [5] in their work.

3. **Refine ChatGPT Approach:** The present approach employed with GPT-3.5-turbo necessitates further refinement. In order to enhance the precision of sarcasm detection within product evaluations, we will investigate more effective prompt designs.

4. **Evaluate GPT-4 Performance:** Leveraging a subset of our dataset, we plan to test the effectiveness of GPT-4 for sarcasm detection, keeping in mind the associated costs.

5. **Interpretation and Rephrasing (Time Permitting):** One of the biggest objectives is to make our model comprehend and reformulate sarcastic or amusing reviews into unambiguous and devoid-of-sarcasm textual content. This approach facilitates the identification of genuine customer feelings, even when they are expressed indirectly through sarcasm or comedy.

6. **Automated Response Generation (Time Permitting):** If feasible within time, our objective is to create a functionality that enables our sarcasm detection model to autonomously generate responses that consider the tone of a user's comment or query.

References

[1] Issa Annamoradnejad and Gohar Zoghi. Colbert: Using bert sentence embedding in parallel neural networks for computational humor, 2020. arXiv:2004.12765, updated Dec 2022. 1

[2] Center for Information and Language Processing (CILP). Mustard++. https://github.com/cfiltnlp/MUStARD_Plus_Plus, 2022. Accessed: Jan 15, 2022. 2

[3] Elena Filatova. Sarcasmamazonreviewscorpus. https://github.com/ef2020/SarcasmAmazonReviewsCorpus, 2012. Paper: Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing, Proceedings of LREC 2012. 2

[4] Aishwarya Gupta, Avik Pal, Bholeshwar Khurana, Lakshay Tyagi, and Ashutosh Modi. Humor@iitk at semeval-2021 task 7: Large language models for quantifying humor and offensiveness. arXiv:2104.00933, accepted at SemEval 2021. 1

[5] Sahil Jain, Ashish Ranjan, and Dipali Baviskar. Sarcasm detection in amazon product reviews. *IJCSIT*, 2018. 1, 2, 4

[6] Rishabh Misra. News-headlines-dataset-for-sarcasm-detection. https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection, 2019. Accessed: Jul 3, 2019. 1

[7] Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, Los Angeles, California, USA, 2016. Sarcasm Corpus V2: https://nlds.soe.ucsc.edu/sarcasm2. 2

[8] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. *COLING*, 2016. arXiv:1610.08815. 1

[9] Hamed Yaghoobian, Hamid R. Arabnia, and Khaled Rasheed. Sarcasm detection: A comparative study, 2021. arXiv:2107.02276. 1

[10] Yftah Ziser, Elad Kravi, and David Carmel. Humor detection in product question answering systems. *Amazon Science*, 2020. Conference: SIGIR 2020. 1