Ronald Chen

Creating A.I. Enabled Systems

NLP Sarcasm Detector

1 Introduction

This offshoot branch of an NLP processing system is an application that enables the user to predict whether written media is sarcastic/satirical or not. Because the tonal aspects of sarcasm/satire are hard to process in its written form, it can be very hard to differentiate from its straightforward counterpart. This application is mainly targeted to individuals who find it difficult to differentiate sarcasm in the written form. This application uses readily available as well as self-identified sarcastic/satirical public media postings to predict the probability of its sarcastic nature. By scraping sites like Reddit, The Onion, Huffington Post, etc. for posts and headlines confidence can be established against selection bias.

There are six key elements to create an AI Enabled system, which includes: Decomposition, Domain Expertise, Data, Design, Diagnosis, and Deployment. These six elements can be referenced as the 6 D's of Creating AI Enabled systems (Figure 1) and it takes a holistic view of creating AI Enabled systems from the initial concept through design and deployment. The following sections will describe each component in greater detail.

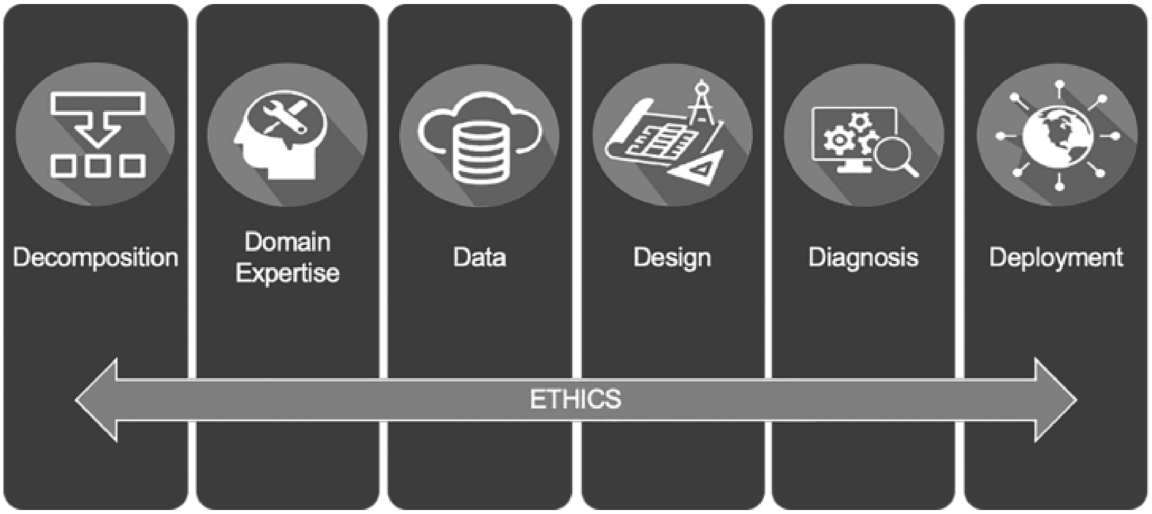


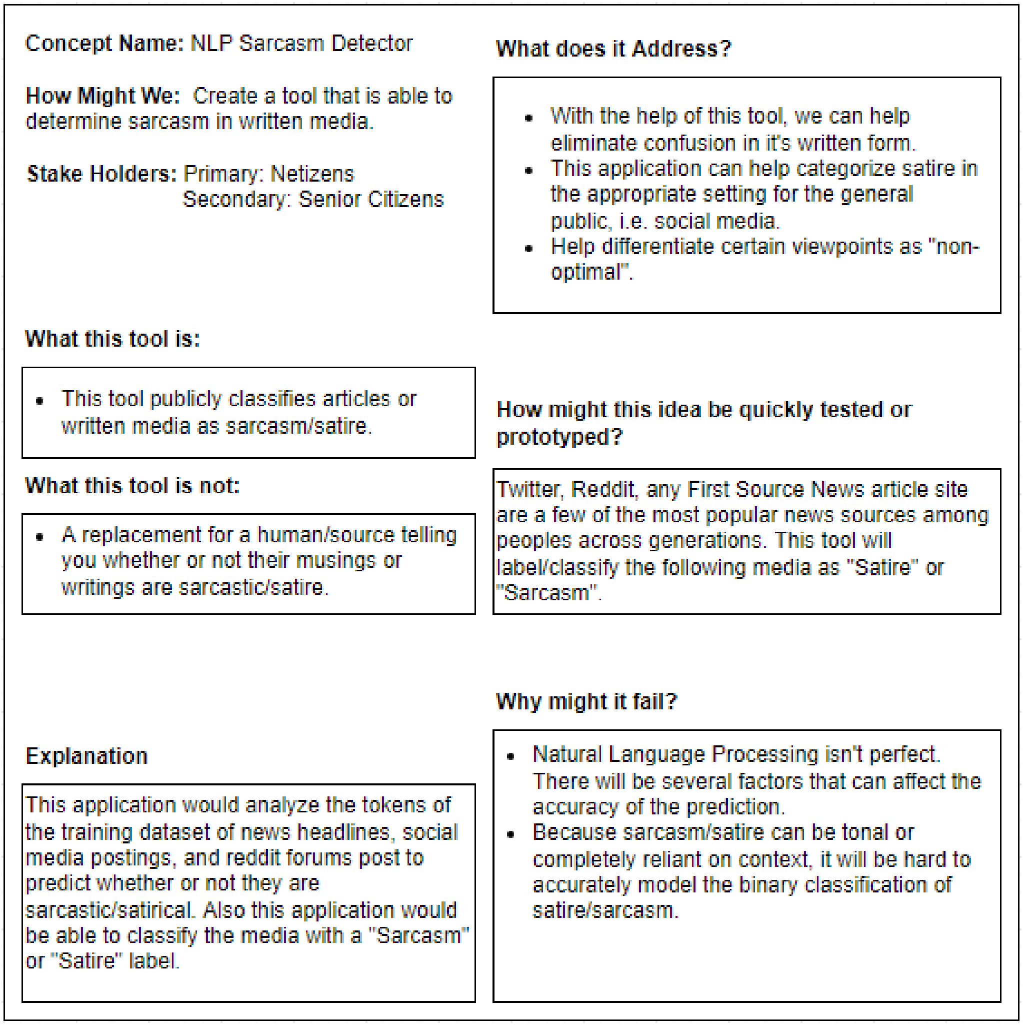
Figure 1: 6D's of Creating AI Enabled Systems

2 Decomposition

"This decomposition component of the 6 D framework mainly consists of refining the technology concept, understanding technology use, and assessing value." – Dr. John Piorkowski, JHU

There aren't many tools online that analyze the writings of written media from a variety of sources. Mainly because it is primarily hard for the average person to read the tone or contextual background of what they're reading from just the excerpt they are looking at. Thus, an interactive application that utilizes open-source models and data scrapings from public platforms can be a good starting point for the average internet user.

2.1 Concept Map



Concept Name: NLP Sarcasm Detector

How Might We: Create a tool that is able to determine sarcasm in written media.

Stake Holders: Primary: Netizens
Secondary: Senior Citizens

What this tool is:

- This tool publicly classifies articles or written media as sarcasm/satire.

What this tool is not:

- A replacement for a human/source telling you whether or not their musings or writings are sarcastic/satire.

Explanation

This application would analyze the tokens of the training dataset of news headlines, social media postings, and reddit forums post to predict whether or not they are sarcastic/satirical. Also this application would be able to classify the media with a "Sarcasm" or "Satire" label.

What does it Address?

- With the help of this tool, we can help eliminate confusion in it's written form.
- This application can help categorize satire in the appropriate setting for the general public, i.e. social media.
- Help differentiate certain viewpoints as "non-optimal".

How might this idea be quickly tested or prototyped?

Twitter, Reddit, any First Source News article site are a few of the most popular news sources among peoples across generations. This tool will label/classify the following media as "Satire" or "Sarcasm".

Why might it fail?

- Natural Language Processing isn't perfect. There will be several factors that can affect the accuracy of the prediction.
- Because sarcasm/satire can be tonal or completely reliant on context, it will be hard to accurately model the binary classification of satire/sarcasm.

Figure 2: Concept Map for NLP Sarcasm Detector

2.2 System Design

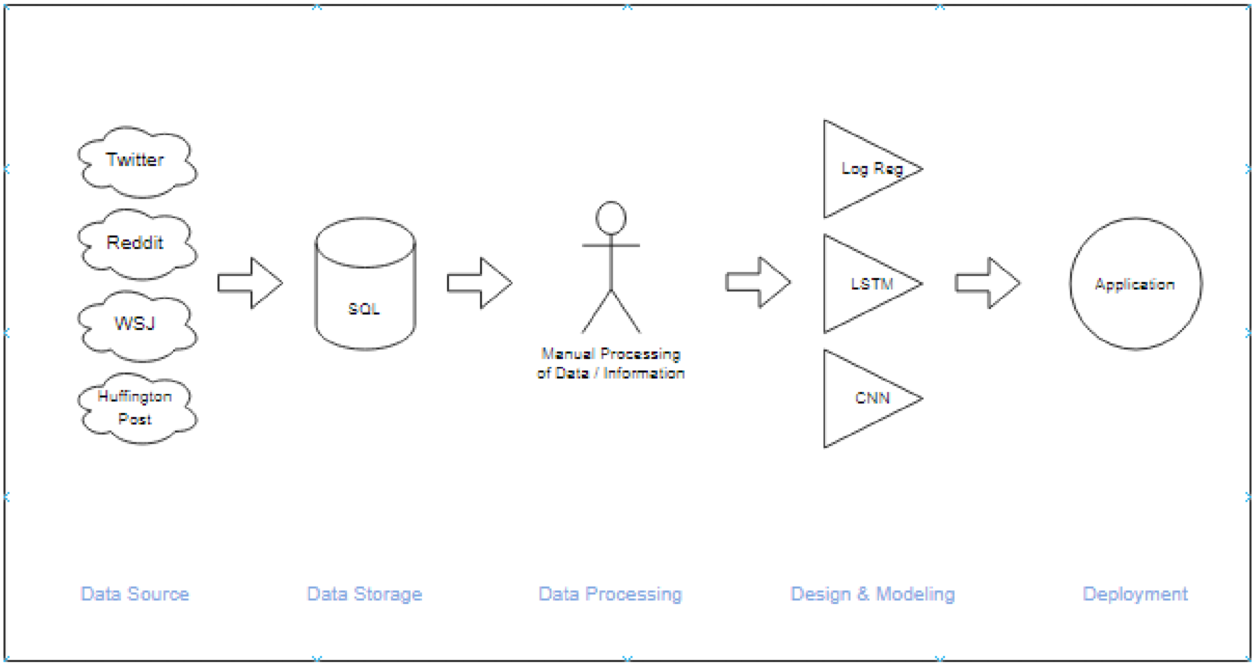


Figure 3: High Level Systems Design

In addition to the binary classification of sarcasm/satire, the data is processed into tokenizer and a word cloud is generated. This demonstrates the frequency of specific words which were categorized via the Models as sarcasm/satire and displays their varying frequency with respect to their font size, i.e. the bigger the word, the more frequent the word appears in sarcastic sentiments. Via experimentation across several deployment strategies, it was most feasible to deploy onto a local client, as the hardware requirements to model any Neural Network based model were more expensive than on IoT.

3 Domain Expertise

"[Domain Expertise] plays an important role beyond decomposition, providing key support during the subsequent phases of the framework. For example, domain experts bring credibility to community members who will be the ultimate users of AI-enabled systems and can assist with adoption of new technology." – Dr. John Piorkowski, JHU

The best way to categorize media as satirical or not is to consult the creator of the media. Because satire is more of an art form rather than science, we can gain a deeper understanding of what typically constitutes satire on existing media for consumers. Furthermore, by consulting these domain experts, inputs and features can be built towards the application. Below are the domain experts to be identified.

3.1 Native Speakers

Native speakers possess a wide repertoire of techniques to convey whatever message they can either speak or write. One such linguistic vehicle native speakers may choose to convey their message is through the art of sarcasm. Native speakers can provide inputs and the definition of what typically constitutes messages on a fundamental level as well as provide a more "artful" indicator or feature.

3.2 News Outlets

News outlets have several ways to draw consumers in, from pure journalism to satirical editorials to draw attention towards a specific point of interest. News outlets typically can be useful as they provide the most "scientific" take on sarcasm as an artform to sell such pieces for other consumers. Sources, such as The Onion or Charlie Hebdo, almost exclusively rely on satire as their main business model, and thus can be extremely useful in making predictions more accurate when compared to other sources, i.e. Wall Street Journal.

3.3 Internet Users

Internet users engage in communication with others on the internet on a daily basis. Typically, these forms of communication can range from slang to stoic communication. Because all media from these internet users are public access and can be aggregated in a much more easily accessible way in comparison to native speakers, internet users can provide important aspects of what constitutes sarcasm in a greater quantity than the other domain experts.

4 Data

"In AI projects, the data engineering phase is generally the most resource intensive, involving collecting, moving, storing, transforming, labeling, and optimizing data to facilitate the design of the system." – Dr. John Piorkowski, JHU

Data is the biggest and the most important aspect of any AI enabled system. There are several techniques to engineer data, one such data engineering perspective is the Extract, Transform, and Load (ETL) process. The Extraction step addresses the collection of raw data using a range of various methods and quality. The transformation is critical and generally is the most resource intensive step. This transformation typically consists of processing data, load and cleaning, feature normalization and extraction, stop words, etc. The pipeline for this transformation process can also be denoted as pre-processing. Lastly, the load step consists of the end-stage of data in which in can be processed by the model.

4.1 Collection and Storage

This application requires data, like any AI project, to build predictive models. The most important data is the data collected from internet users for sarcastic messages and owner classified labels. The data includes, comment, author, subreddit/topic, score (upvotes or retweets/likes), date/time, and parent comment/tweet. There are many ways to locate and extract this data from the internet. Websites such as Twitter and Reddit provide real time and historical data for free. Figure 4 shows a word cloud of the most commonly used words in sarcastic comments. Figure 5 shows a word cloud of the most commonly used words in straightforward comments.

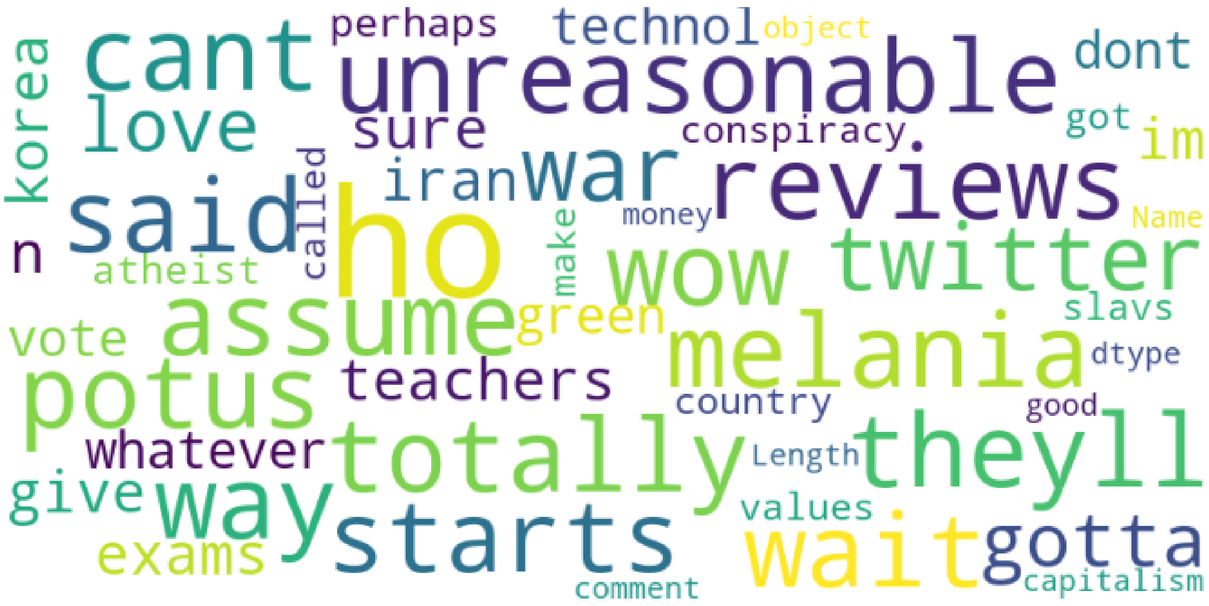


Figure 4: Sarcastic Word Cloud

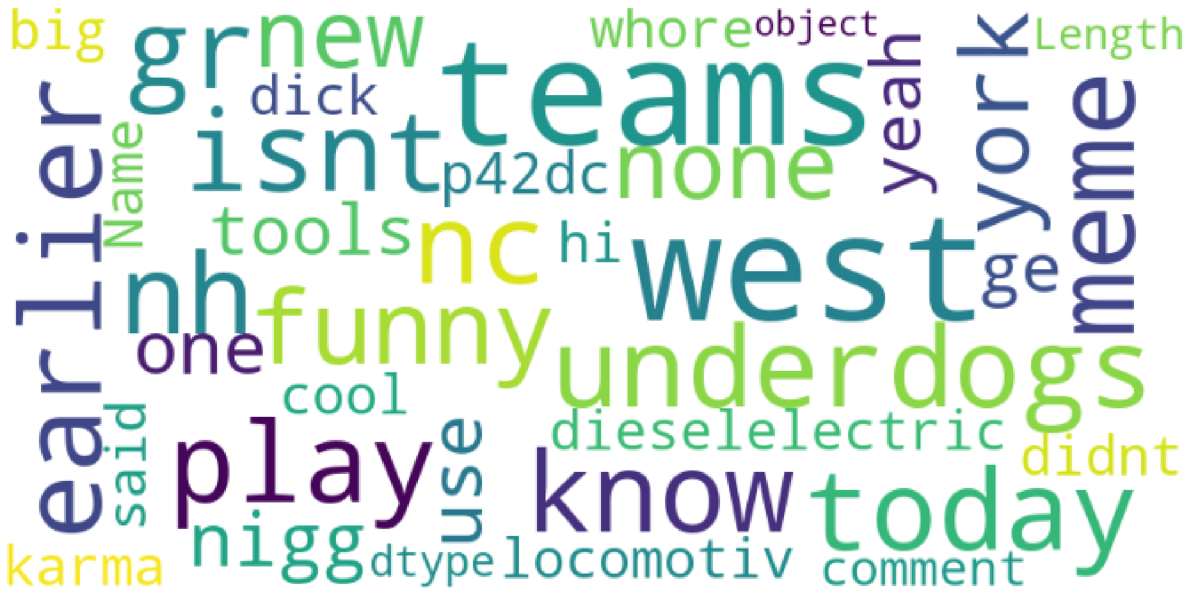


Figure 5: Non-sarcastic Word Cloud

To store this data, MongoDB was used. Due to the nature of the "big data era" a non-relational data storage strategy was used because of the exponential nature of the quantity of data in the internet. In particular, this makes a NoSQL strategy, e.g. MongoDB, an ideal candidate for this application.

4.2 Processing Data

Classification of sarcasm depends on many factors like, mood, time, information, context, landscape, etc. Many of these are captured in the features in the dataset. These features are processed and tokenized and tf-idf vectored to better classify their nature.

Pre-processing Data

Several requirements were necessary to process the data in a way to make it more feasible to be processed by the model. Because of the differencing in values of ASCII characters, each character must be normalized and thus was converted to an all-lowercase feature. Additionally, all characters except for alphabetic and whitespace characters were captured.

Feature Extraction

Several factors were considered in what constitutes a sarcastic comment. Figure 6 shows the average length of a sarcastic comment's length in comparison to its parent comment.

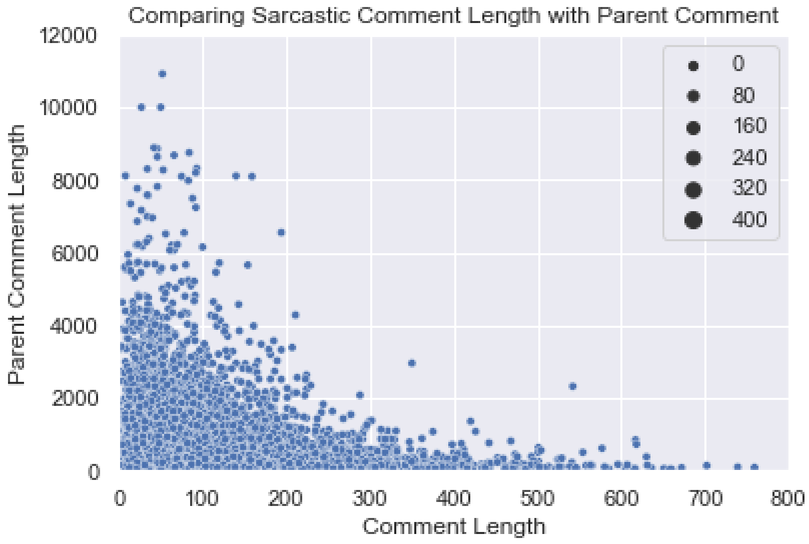


Figure 6: Comparison of Lengths from Sarcastic Child Comment to Neutral Parent

As seen above, sarcastic comments tend to be much shorter in length in comparison to their parental counterpart. This can be due to several factors ranging from the hypothetical requirement to be short and witty to difficulty in being creative. Figure 7 below shows the count of sarcastic comments per day across different weekdays.

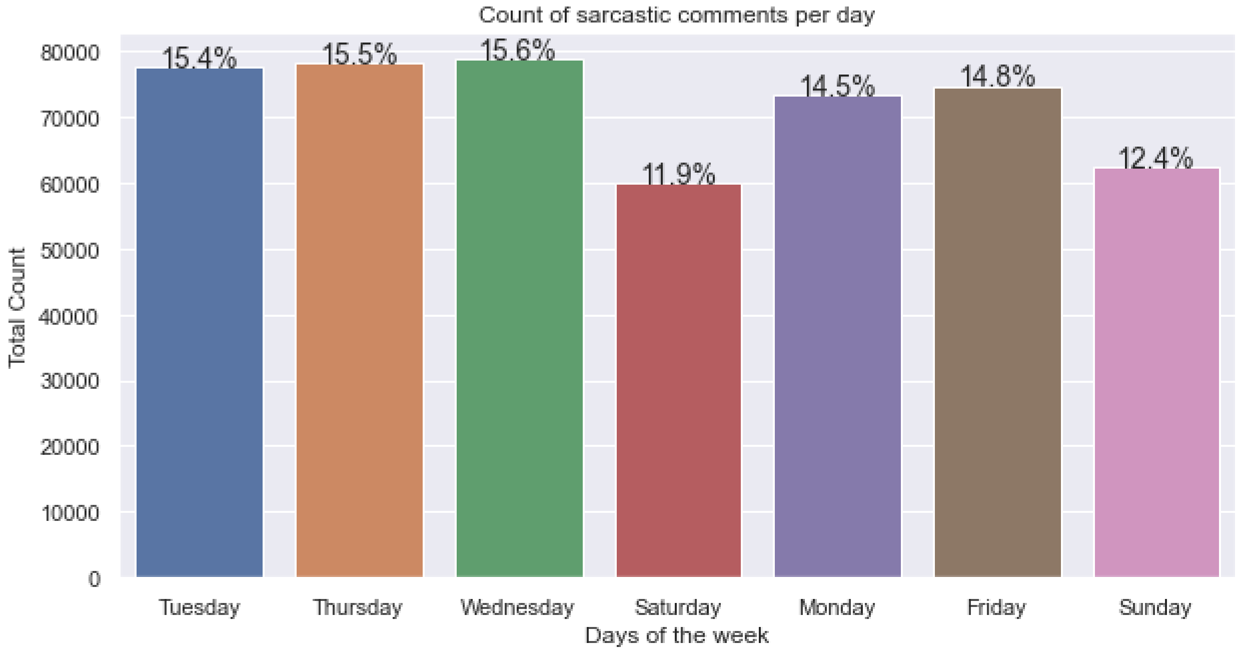


Figure 7: Count of Sarcastic Comments on Different Weekdays

As seen above, most sarcastic comments are typically made during weekdays, i.e. Monday to Friday. Figure 8, shown below, shows the count of total comments in respect to the most commented topics / categories.

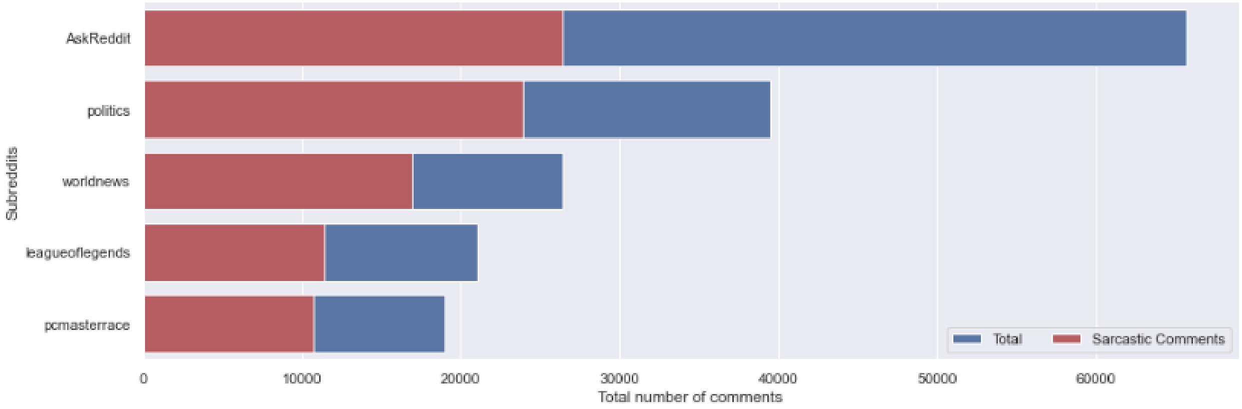


Figure 8: Total Count of Neutral and Sarcastic Comments

As shown above, certain topics and subreddits have a greater likelihood of being sarcastic in comparison to their neutral counterparts. The only subreddit that has more neutral comments is the one that typically requires a more "serious" response. Every other subreddit has more sarcastic comments than neutral comments.

5 Design

"The AI field is characterized by several algorithmic approaches ranging from rule-based systems to machine learning algorithms. In practice, many AI-enabled solutions may include a combination of algorithm classes." – Dr. John Piorkowski, JHU

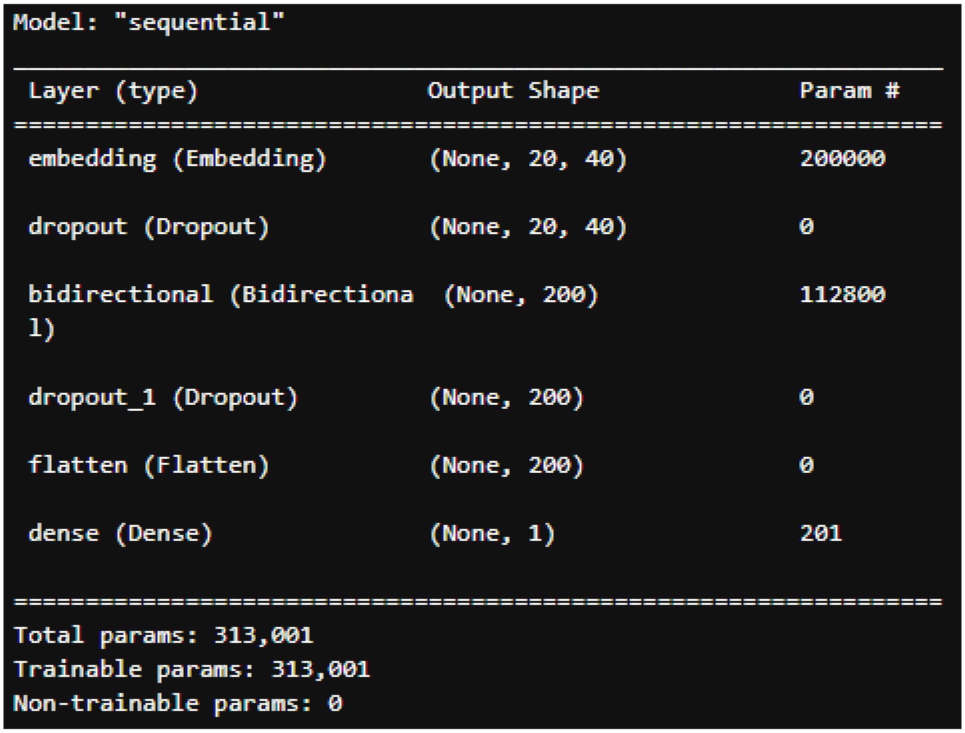
Once the dataset has been processed, the next step is to build the model. For demonstrative and academic purposes, this application has multiple models: Logistic Regression, Long Short Term Memory, and Convolutional Neural Network. This pipeline starts builds these models, evaluates it on the evaluation dataset, and tunes the parameters to test the final performance of each model of the test dataset. Due to the hardware intensive nature of the LSTM and CNN models, it was not feasible to have this model be executed on the IoT. While the data can be stored in an non-relational database, the neural networks required much more hardware than was allotted, and would thus crash.

5.1 Logistic Regression

In this application, the default and control model used was the Logistic Regression Model. Logistic Regression is very useful as it predicts a classification label in a very straightforward manner. Additionally, due to its nature, it is much faster in comparison to other models, allowing a Logistic Regression application to be hosted on the IoT.

5.2 Long Short Term Memory Network

An LSTM network is very powerful in sequence prediction problems because they are able to store past information. In this way, it allows the model to provide contextual information past a sequence of characters in the data. This is important because context is very important in sarcasm prediction because context can be crucial in sarcasm. Figure 9 shows the LSTM model architecture.

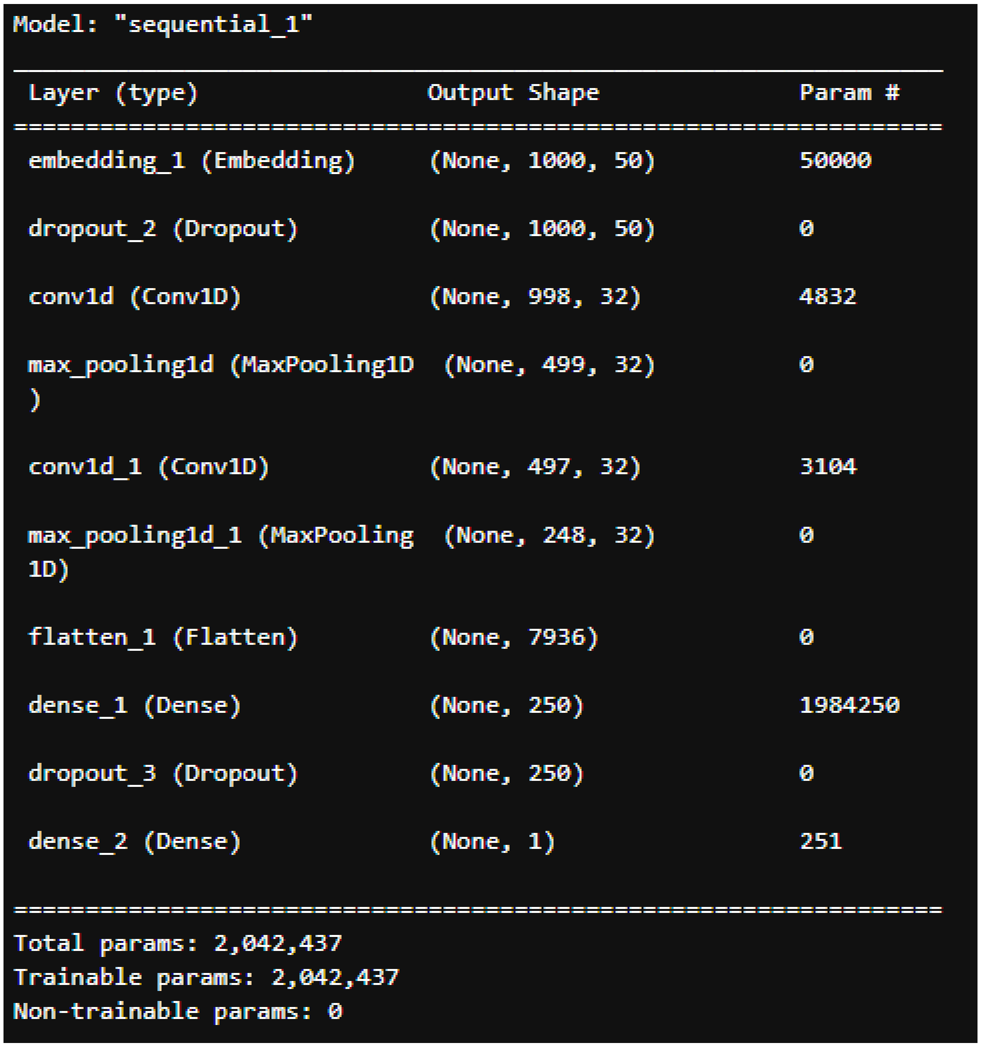


```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 20, 40)            200000

 dropout (Dropout)           (None, 20, 40)            0

 bidirectional (Bidirectiona  (None, 200)              112800
 l)

 dropout_1 (Dropout)         (None, 200)               0

 flatten (Flatten)           (None, 200)               0

 dense (Dense)               (None, 1)                 201

=================================================================
Total params: 313,001
Trainable params: 313,001
Non-trainable params: 0
```

Figure 9: LSTM Model Architecture

5.3 Convolutional Neural Network

This neural network is powerful because it can detect features in an unsupervised fashion. As opposed to the Logistic Regression model or the LSTM model, this model can learn distinctive features on its own.



```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 1000, 50)          50000

 dropout_2 (Dropout)         (None, 1000, 50)          0

 conv1d (Conv1D)             (None, 998, 32)           4832

 max_pooling1d (MaxPooling1D  (None, 499, 32)          0
 )

 conv1d_1 (Conv1D)           (None, 497, 32)           3104

 max_pooling1d_1 (MaxPooling  (None, 248, 32)          0
 1D)

 flatten_1 (Flatten)         (None, 7936)              0

 dense_1 (Dense)             (None, 250)               1984250

 dropout_3 (Dropout)         (None, 250)               0

 dense_2 (Dense)             (None, 1)                 251

=================================================================
Total params: 2,042,437
Trainable params: 2,042,437
Non-trainable params: 0
```

Figure 10: CNN Model Architecture

5.4 Hyper-Parameterization

Hyperparameters are used to control the learning process of the model. To optimize the above models, the following hyper parameters were identified and tuned:

1. Batch Size
2. Number of Epochs
3. Number of Layers
4. Number of Nodes
5. Activation Functions
6. Learning rate & Decay

6 Diagnosis

"The diagnosis component addresses how AI-enabled systems are assessed and what metrics are used. The design section identified four algorithm classes that bring a different set of metrics. Typical metrics for supervised machine learning algorithms include accuracy, a confusion matrix, per class accuracy, log loss, precision, recall, mean average precision, and Area Under the Curve (AUC)." – Dr. John Piorkowski, JHU

Once the models have been trained and tested, the next step is to measure the performance of the models.

6.1 Evaluation Metrics

The main evaluation metrics used are identified below to evaluate the performance of the model on the test dataset.

Logistic Regression

F1: 69.37%

Long Short Term Memory

F1: 71.88%

MSE: 0.5638

MAE: 0.7053

Convolutional Neural Network

F1: 67.22%

MSE: 0.5957

MAE: 0.6746

6.1.1 F-Score (F1)

The F-score is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have identified as positive.

6.1.2 Mean Squared Error (MSE)

The mean squared error is an estimator that measures the average of the squares of the errors. The average squared difference between the estimated values and the actual value.

6.1.3 Mean Absolute Error (MAE)

Mean absolute error is a metric which finds the average absolute distance between the predicted and target values.

6.2 Bias Issue

Data bias can be observed in the reddit comments which are used to determine sentiment scores. More popular opinions were weighted more popularly as they were more confidently sarcastic. However, due to the nature of "popularity" the bias of political current events is introduced. The popular comments far outweighed the neutral or negative comments, which make the data positively-biased.

To overcome this bias, data was scraped from multiple sources across different platforms and communities. By aggregating different sources, this positive bias could be mitigated and analyzed for the reason this bias exists.

6.3 Ethical Issues

Logistic Regression is prone to biases that exist within the data. Because Logistic Regression tends to be a binary classification, nuances in the result can not be represented. Additionally, it is wholly dependent on the dataset. It is trivial to explain the classification of a Logistic Regression model. In an LSTM model, it becomes more difficult to explain the classification behind the prediction. Making an LSTM model being more ethical than a Logistic Regression Model. However, the most biased model would be the Convolutional Neural Network. The CNN is wholly dependent on the data that is being fed into the model. Because it is an unsupervised learning model, it will pick the most "optimal" features which can be completely affected by the biased data.

Accountability in the auditing process would help consumers better understand, trust, and manage these AI enabled systems. How transparent the model's classification process and how trivial the explanations are would allow greater trust in the application.

7 Deployment

"There is no one-size-fits-all approach for AI-enabled system deployment, but it is important to note that AI technologies are generally employed in a broader system context. With the prevalence of cloud computing and client/server models, AI-enabled systems leverage this type of computing paradigm. " – Dr. John Piorkowski, JHU

The last D of the 6D framework is deployment. This final phase is where the application is deployed and continuously monitored to ensure performance. The strategy for this application is a Hybrid approach, i.e. a combination of server-side and client-side approaches. Where the data and storage is all on a server and the application resides on the client as these clients typically hold more power than a cloud. Additionally, this approach provides the ideal amount of flexibility in which to deploy a more tuned system, depending on the model the user would choose.

By continually monitoring the deployment, the analytics of the applications will be continually observed. As the data grows due to the ever-increasing evolution of the internet, we can introduce more data into the pipeline and further train the model to improve performance

8 Conclusion

With the data aggregated from Reddit and Twitter, in combination with several machine learning models, we can, to a certain degree, develop a system that can differentiate sarcasm from neutral language, up to 75%. However, from what I have learned from this course and this project, is that to build an effective application, the 6D framework is essential. Without properly understanding Decomposition, Domain Expertise, Data, Design, Diagnosis, and Deployment the application could have suffered and may have been much harder to explain the results. Not only does the 6D allow for a smoother build, but it also makes the explanation easier to digest if you can trace the pipeline at each step of the build. With careful consideration of each “D” in the framework, we can modify the application in a modularized manner to appropriate the system for different applications.

Bibliography

Hurley, D. (2008, June 3). The science of sarcasm (not that you care). The New York Times. Retrieved May 8, 2022, from https://www.nytimes.com/2008/06/03/health/research/03sarc.html?em&amp;ex=1213848000&amp;en=79518c9f61e51946&amp;ei=5087%0A

Laura Sebastia Technical University of Valencia, Sebastia, L., Valencia, T. U. of, Eva Onaindia Technical University of Valencia, Onaindia, E., Eliseo Marzal Technical University of Valencia, Marzal, E., &amp; Metrics, O. M. V. A. (2006, January 1). Decomposition of planning problems. AI Communications. Retrieved May 8, 2022, from https://dl.acm.org/doi/10.5555/1143139.1143143

Piorkowski, J. May 8, 2022. The 6-Ds of Creating AI-Enabled Systems. Johns Hopkins University Applied Physics Laboratory

Yin, H. (2020, August 24). 1 - the importance of domain knowledge. Machine Learning Blog | ML@CMU | Carnegie Mellon University. Retrieved May 8, 2022, from https://blog.ml.cmu.edu/2020/08/31/1-domain-knowledge/