Is that Slang?

By Dhaniya Venkat

Everyday language rapidly changes as the meanings of words are contextualized to fit certain cultural or demographic needs. There are also words that have double meanings which, depending on the context, can determine whether a sentence has Gen-Z slang in it or not. For example, 'The stars lit up the sky.' and 'This party is lit!' both use the word 'lit' but only the second sentence is considered Gen-Z slang. This can be very confusing for people of the older generations to understand as some sentences contain slang that does not make sense. How can we better understand GenZ slang that is being used today? In order to tackle this challenge, I decided to make a NLP model that processes sentences and determines whether or not they contain Gen-Z slang. Using Logistic Regression in particular, I was able to create a model that processes inputted sentences as well as went over methods to checking accuracy.

While scrolling through social media or out in public, you may have seen or heard words such as "slay" or "rizz". That raises the question: Where did those words come from? Why are they so popular? These words came to life with the help of the internet. Internet trends spread these words which have now become associated with daily language. Regardless of whether you heard it from a young group of Gen-Zers at the mall or on a social media post, you may or may not have understood what the word means. This is where my model comes in; this problem works with supervised data since we are classifying the sentences as containing Gen Z slang or not. The data is language-based as it contains various types of sentences but also has a categorical aspect. Once you input a sentence into my model, it will tell you whether or not it has Gen-Z slang.

I went through various technical and non-technical articles that helped me gather more information about Gen Z slang and its importance in today's world. The article "The origin story of Gen Z slang" by Hanna Seariac talks about the origins of Gen Z slang and how it is affected by the digital world. For example, the text states, "Since TikTok can lead to videos quickly going viral, it makes sense that slang would evolve more quickly." It also explained how slang helps convey messages with short responses or abbreviations. This seems efficient, however, it is not professional. Slang cannot be used in every situation.

Another article, "The Quad: The evolution of Gen Z slang words and their modern meanings " by Kanishka Mehra talked about how certain words in various contexts could convey a different meaning. The text states, "The word 'lit,' for example, is commonly used nowadays to describe things that are exciting, lively or intoxicating, rather than being literally caught on fire." This also explains the importance of context; words have different meanings depending on how you use them. This is fundamental to languages and it does not apply any less to Gen Z slang.

Lastly, the article "Digital Culture and Social Media Slang of Gen Z " by Eliza M. Jeresano and Marigrace D. Carretero thoroughly researches Gen Z slang among students and its controversy in school. They collected text messages from high school students and categorized the types of slang they found. They also got the opinions of teachers to see whether or not the use of Gen Z slang in schools was accepted. Not only did this article find commonalities in Gen Z slang, it also provided both positive and negative opinions on its usage. Types of Gen Z slang and the different meanings of words are something I take into strong consideration as it is likely to increase or decrease the accuracy score of the model.

I created a language-based data set filled with various types of sentences (N = 508). Half of the dataset contained GenZ slang while the other half did not. Some examples of Gen-Z slang in the dataset were, "You're acting pretty sus." and "He yeeted that pencil across the classroom!" Examples of regular sentences include “The fragrance of lavender lingered in the air.” and “Pink horses galloped across the sea.” I made sure to use a wide variety of sentences for each section so that the model would be accustomed enough to tell the difference between any sentence it is given. The dataset has two columns–one column contains the sentences while another column states whether or not they have Gen-Z slang or not (Yes or No).

Sentence	Contains Gen-Z slang?
"Your opinion is based."	YES
"Those girls over there look so basic."	YES
"I'm almost ready for my date night, just have to beat my face real quick."	YES
"You are my number one bestie."	YES
"You want to get ice cream after school? Bet."	YES
"Wow, when he failed that stunt, that was a big yikes."	YES
I trust everything that's written in purple ink.	NO
The blinking lights of the antenna tower came into focus just as I heard a loud snap.	NO
Siri became confused when we reused to follow her directions.	NO
He found his art never progressed when he literally used his sweat and tears.	NO
"He dabbed on the haters."	YES

Table 1. Dataset sentences and classifications (Gen Z slang or not)

Ultimately, the computer doesn’t understand or process human language like we do; that is why they need to be taught using binary numbers. We can start converting sentences with the help of tokenization. Tokenization helps split sentences into an individual list of words (gets rid of capitalization, punctuation, etc.). Next, we remove stopwords, words that do not help differentiate sentences from each other (or in this case, sentences with Gen Z slang versus regular sentences). Finally, we convert all the useful words into number embeddings called vectors such that each word has its own vector. All the above steps are conducted using a

function called Count Vectorization which tokenizes, removes stopwords, and converts words into vectors. Using the Natural Language ToolKit (NLTK), I was able to remove all the stopwords all well as graph out the most commonly used words:

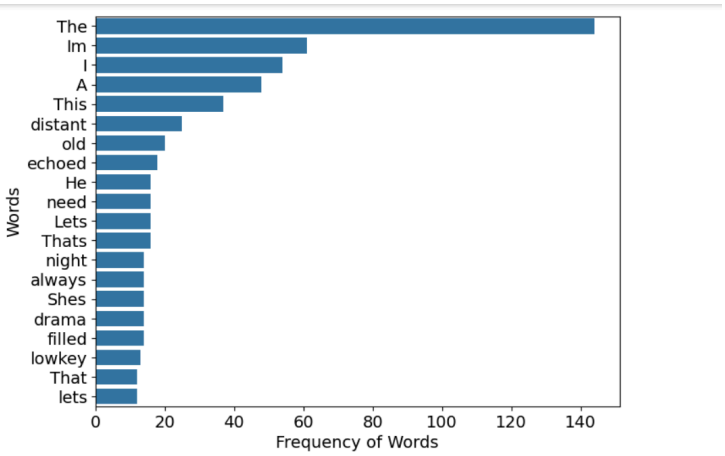
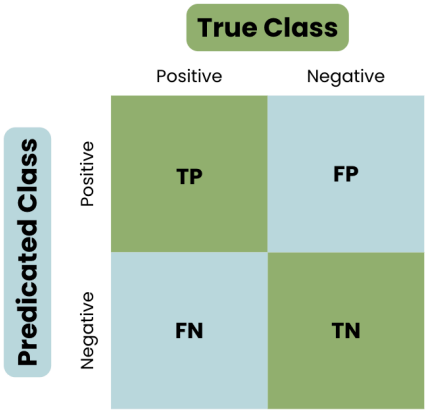


Figure 1. Graph of the most used words found in the sentences of the dataset

Using the dataset, I processed the sentences such that the NLP model would be able to differentiate between the ones containing Gen Z and the regular sentences. I used three methods to calculate the accuracy score: Logistic Regression, Confusion Matrix, and the ROC Curve. Using Logistic Regression, we split our data into training data and testing data (X_train/X_test represents the sentences while y_train/y_test represents whether or not it is slang–Yes or No). The training data is used to train the model to differentiate between Gen Z slang and regular words. The testing data is used to test the accuracy of the model's prediction (based on the training data) compared to the real answer. Since we only need a small portion of our data to be tested, we randomly assigned 20% of the data as testing data. After splitting our data, we create a model and fit it using the training data. Next, we set our prediction using the X_test data. The accuracy score is then calculated based on the predicted result and the actual result.

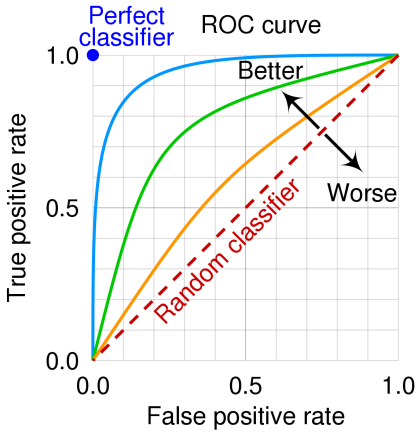
Using the Confusion Matrix, is a way of comparing predictions based on training data with the real answers from the testing data. The predictions are split into four categories: true positives, false positives, true negatives, and false negatives. True positives and true negatives are predictions that are correct while false positives and false negatives are not. Positives and negatives are represented as numbers where 1 is positive and 0 is negative. There are only two numbers because the problem is a classification problem where either a sentence has Gen Z slang or not. We can calculate the accuracy score using the confusion matrix equation: (TP + TN) / (TP + TN + FP + FN). We can also find values such as precision (identifies how accurately the model predicts positives) by using the equation TP / (TP + FP). We can also calculate the recall (identifies how much the model predicts true positives) using the equation TP / (TP + FP).



datacamp.com

Figure 2. Visual representation of the Confusion Matrix and how predictions are classified

Using the ROC Curve is a more visual way of seeing how accurate your model is. It plots the true positive rate against the false positive rate (the amount of sentences that do have Gen Z slang as compared to the sentences that were thought to have Gen Z slang. The closer the curve is toward 1 on the true positive axis, the more accurate the model.



medium.com

Figure 3. Graph of accuracy for ROC Curve

The Logistic Regression model has an accuracy score of 94.12%. The Confusion Matrix also had an accuracy score of 94.12%, a recall of 90.91%, and a precision of 90.04%. Setting the max features to 800, which picks out 800 of the most common words in my preprocessed data set, helped differentiate the different types of sentences better. The ROC Curve displayed a curve that was close to 1 on the true positive axis which shows that the model is pretty accurate.

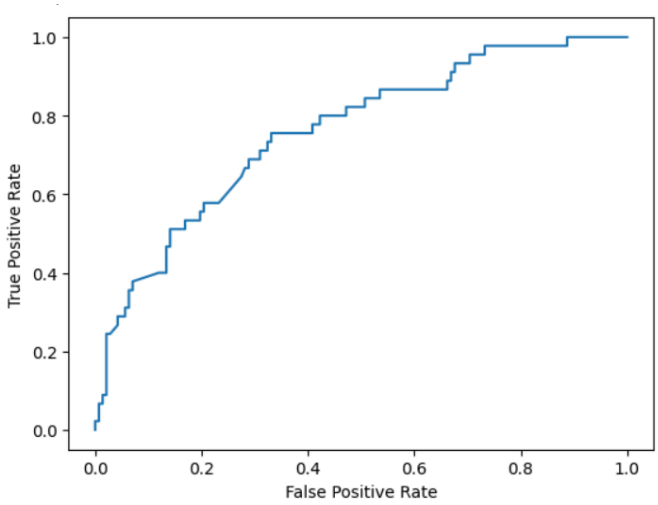


Figure 4. ROC Curve of the Gen Z slang dataset

Using the Logistic Regression model I trained earlier, I incorporated it into a program that uses user input and determines whether the user input uses Gen-Z slang or not. First, I converted the user input into a list. Next, I turned it into vectors using the CountVectorization function. I then used the logistic model I trained earlier to predict whether or not the user input had Gen Z slang. This model worked pretty well with determining the difference between sentences like 'The stars lit up the sky.' and 'This party is lit!' A problem with the model is that it has difficulty differentiating short sentences like "Hi there" or "What's up?". The model believes these sentences contain Gen Z slang even though that's not true. A possible reason for this issue is that my dataset didn't contain enough short sentences for the model to be trained on. Another reason for this could be that I didn't add many non-Gen Z greetings to the dataset. As a result, the model struggles with such sentences, yet overall, does a nice job.

Throughout the research process, I focused on understanding the relevance of Gen-Z, how it came to be, and how to distinguish it using NLP. Gen-Z slang is a more recent development that evolved along with social media. Since it is different from what we consider to be the "normal" English language, I wanted to create a model that helps differentiate Gen-Z sentences from normal sentences. Using different methodologies such as Logistic Regression, the Confusion Matrix, and the ROC Curve helped us better understand the accuracy of the model and how we can better it. When applying the model using user input it did pretty well in telling the difference between Gen-Z slang and regular sentences, but got confused with shorter sentences. This was most likely due to the data consisting of mostly longer sentences (for the normal sentences), so it wasn't trained much on shorter sentences.

My next steps for the model would be adding sentences of different lengths to the training data as well as attempting to add definitions of Gen-Z slang words for sentences with Gen-Z slang so that if the user input contains Gen-Z slang, it will also provide the user an understanding of what it means. Overall, this model can be used to distinguish many other types of sentences from regular English such as Shakespearean English or even slang from the 90s! There are so many possibilities for this project and so many ways we can use NLP!

I would like to thank my mom for the encouragement, my mentor, Nancy Zhu, for the wonderful help and support throughout my project, and the Inspirit AI program for providing resources and giving me the opportunity to complete such a project!

References

7ESL7ESL proudly offers an exceptional English-learning experience through our app. (2023, November 4). *The 100 most common Gen z slang words • 7esl*. 7ESL. https://7esl.com/gen-z-slang/

Jeresano, E. M., & Carretero, M. D. (2022). Digital culture and social media slang of Gen z - uijrt. https://uijrt.com/articles/v3/i4/UIJRTV3I40002.pdf

Mehra, K. (2019, October 24). *The Quad: The Evolution of Gen Z slang words and their modern meanings*. Daily Bruin. https://dailybruin.com/2019/10/24/the-quad-the-evolution-of-gen-z-slang-words-and-their-modern-meanings

Random sentence generator - 1000+ random sentences. Random Word Generator. (n.d.). https://randomwordgenerator.com/sentence.php

Seariac, H. (2023, February 13). *The origin story of Gen z slang*. Deseret News. https://www.deseret.com/2023/2/13/23574658/gen-z-slang

ChatGPT (FOR SENTENCES ONLY)

Wikimedia Foundation. (2024, January 18). *List of generation Z slang*. Wikipedia. https://en.wikipedia.org/wiki/List_of_Generation_Z_slang (FOR SENTENCES ONLY)

Huilgol, P., & Shah, R. (n.d.). Precision and Recall | Essential Metrics for Machine Learning (2024 Update). Analytics Vidhya. Retrieved February 13, 2024, from https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/