Analyzing Tinder Data with NLP to Predict Digital Dating Success

(CS5785) Applied Machine Learning, Cornell Tech, Fall 2023

Jae Kim
Master's student, Health Tech
Cornell Tech
jk2765@cornell.edu

Yoseph Yan
Master's student, Electrical and Computer Engineering
Cornell Tech
ypy4@cornell.edu

ABSTRACT

This study utilizes natural language processing and machine learning to analyze initial messages on digital dating platforms, specifically Tinder. The goal is to identify key features in these messages that correlate with digital dating success, defined as longer conversation lengths. We apply four distinct machine learning models - Logistic Regression, Naive Bayes Classifier, Neural Network using Multilayer Perceptron, and Random Forests - to predict the likelihood of successful interactions. This research offers insights into the dynamics of digital dating communication and potential strategies to enhance user experiences on these platforms.

KEYWORDS

machine learning, tinder, textual data mining, Logistic Regression, Natural Language Processing

ACM Reference Format:

Kim, J., & Yan, Y. 2023. Analyzing Tinder Data with NLP to Predict Digital Dating Success: Cornell Tech, Applied Machine Learning (CS5785), Fall 2023, New York, NY, USA, 5 Pages

1 INTRODUCTION

This project explores the art of the initial message in online dating on Tinder, the leading dating app with over 70 million users. As the digital arena for romance, Tinder has become a cultural mainstay for singles seeking connection. Despite its popularity, success on the platform varies widely among users. Our research delves into Tinder's extensive conversation data to discern the words that sustain engagement, aiming to decode the elements of successful communication. By analyzing interactions that surpass the average conversation length, we provide insights to guide users in crafting messages that resonate and engage, leveraging the powerhouse of digital dating that Tinder has become.

2 BACKGROUND

2.1 Dataset

The study utilizes a comprehensive dataset sourced from Swipestats.io, an online platform that consolidates usage data from the Tinder application. By engaging with the platform's curator, Kris, we secured access to a substantial, anonymized dataset that captures a multifaceted picture of Tinder user activity. This dataset includes diverse elements ranging from user demographics and preferences to detailed profiles and rich conversation histories. It offers a robust basis for an in-depth exploration and analysis of patterns and trends in digital dating behaviors.

2.2 Preprocessing

The dataset comprised data from 1,209 Tinder users, encapsulating 23,269 distinct features. Initial preprocessing involved the extraction of critical attributes, including user IDs, message content, match IDs, and average conversation lengths. We prioritized the first message in each conversation by eliminating subsequent duplicates, focusing on the initial interaction. Advanced text processing techniques such as stop word removal, tokenization, and lemmatization were employed. Additionally, initial messages were categorized into various types like questions, GIFs, and greetings to further refine our analysis. This rigorous preprocessing yielded a dataset that offers a granular perspective on the initial messaging patterns of Tinder users.

2.2 Data Visualization

Using the dataset that we preprocessed, we plotted the average conversation length for each message type. Figure 1 shows that GIFs resulted in the shortest average conversation length and greetings resulted in the highest average conversation length. This tells us that a user may want to refrain from sending GIFs in order to maintain a more engaging conversation. Figure 2 shows the number of times each message type was said to a match. By far, greetings were the most used initial message. Followed by a question and then a GIF.

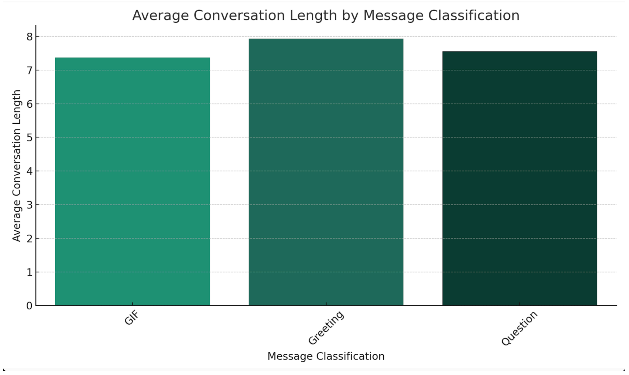

Figure 1: Message Type vs Average Conversation Length

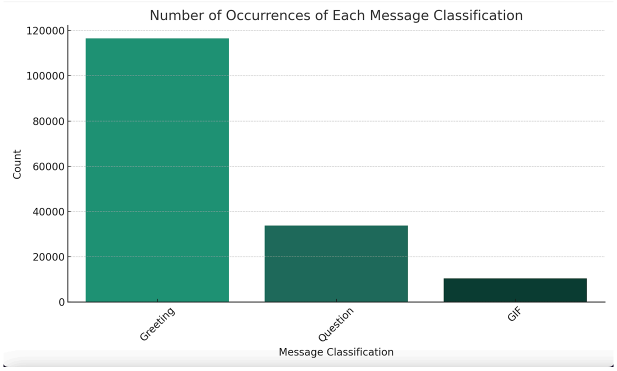

Figure 2: Count of Message Types

3 METHOD

This study employs four distinct machine learning models for classification tasks, specifically chosen for their suitability in analyzing digital dating interactions. These models are Logistic Regression, Naive Bayes Classifier, Random Forests, and Neural Network using Multilayer Perceptron.

3.1 Logistic Regression

As a fundamental binary classification algorithm, logistic regression is utilized for its effectiveness in deciphering binary outcomes, such as predicting whether a message will lead to a continued conversation. Its adeptness in processing high-dimensional, sparse text data is further refined with the introduction of a regularization term to ensure model generalizability across varied conversational scenarios.

3.2 Naive Bayes Classifier

The Naive Bayes Classifier, known for its simplicity and efficiency with textual data, is applied to predict the likelihood of sustained interactions. This model assumes feature independence and is well-suited to our dataset's dimensional complexity.

3.3 Random Forests

Incorporating the Random Forest algorithm, we address the challenge of complex, non-linear data relationships typical in digital dating predictions. This ensemble learning approach amalgamates multiple decision trees, aiming for heightened accuracy and reduced overfitting risk, thereby providing a comprehensive model evaluation in comparison to logistic regression and Naive Bayes.

3.4 Neural Network using Multilayer Perceptron

The Multilayer Perceptron (MLP) Neural Network is the final model in our methodological arsenal. Chosen for its capability to identify intricate, non-linear patterns, MLP is instrumental in exploring the depths of NLP within the context of digital dating, offering comparative insights against the aforementioned models.

4 EXPERIMENTAL ANALYSIS

4.1 Initial Observations

In the preliminary phase, our analysis meticulously parsed the dataset to establish a correlation between specific lexicon usage in initial messages and the subsequent conversation length. Through a comprehensive linguistic analysis of over 1209 Tinder user exchanges, we discerned a pattern where emotionally positive words such as “love”, “haha”, and affirmations like “yes” were indicative of longer conversations. Conversely, commonplace greetings like “hi” and “hey”, despite their high frequency, were often precursors to shorter interactions. Intriguingly, references to pets, particularly “dog” and “cat”, were frequently present in more engaged dialogues, underscoring the social value of pet ownership in digital dating dynamics. These observations form the basis for our hypothesis that the initial message's content is a strong predictor of conversation longevity, warranting a deeper exploratory analysis through machine learning models.

4.2 Setup

The dataset was prudently divided into training, development, and test subsets to facilitate a comprehensive evaluation of the predictive models. A bag-of-words model was then applied to distill the corpus, focusing on terms appearing in a minimum of three instances. This methodology refined our feature set to 17,199 words from an initial lexicon of 66,329, ensuring a balance between model complexity and computational efficiency.

4.3 Initial Training

A proportionate split of 60% training, 20% development, and 20% test subsets ensured a rigorous training and validation process for our suite of machine learning models. The precision metrics across the training and development sets indicated a well-calibrated fit, negating the presence of underfitting or overfitting phenomena within our computational framework.

Table 1: Prediction Accuracy with Threshold of 3

	Logistic	Naive Bayes	MLP	Random Forest
Train Precision	79.19%	75.84%	77.57%	76.52%
Dev Precision	74.82%	73.39%	75.49%	74.63%

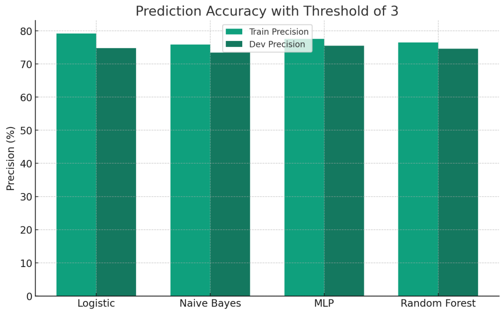

Figure 3: Prediction Accuracy with Threshold of 3

4.4 Training with Threshold of 5

For this experiment, we increase the bag of words threshold from 3 to 5. This is to reduce the number of features to see if it has any effect on the prediction accuracy. It is also to train our model using words that are more common. After training the models, we concluded that it does not have much of an impact on the prediction accuracy. The prediction accuracy does not change much with a threshold of 5.

Table 2: Prediction Accuracy with Threshold of 5

	Logistic	Naive Bayes	MLP	Random Forest
Train Precision	78.2%	75.07%	77.97%	76.89%
Dev Precision	74.64%	73.22%	75.59%	74.28%

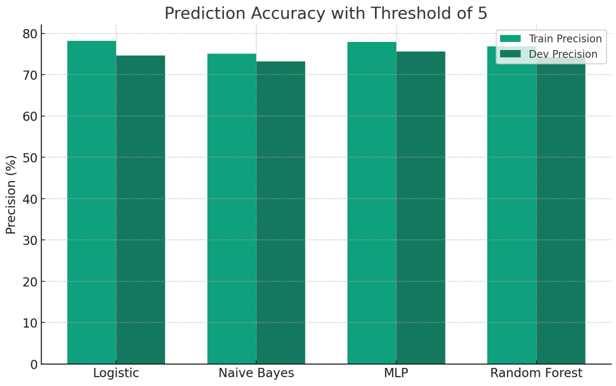

Figure 4: Prediction Accuracy with Threshold of 5

4.5 Testing

Upon a comprehensive evaluation of various models, the Logistic Regression emerged as the most accurate. Consequently, we proceeded with this model for a more detailed performance analysis. The testing phase yielded a prediction accuracy of 59.24%. This result signifies that the Logistic Regression model successfully predicted whether an initial message would lead to a longer-than-average conversation with a reliability of 59.24%. Such a finding underscores the model's capability to effectively discern key elements in initial messages that are likely to foster sustained engagement on the platform.

Table 3: Prediction Accuracy of Logistic Regression

	Training Precision	Test Precision
Logistic Regression	74.62%	59.24%

5 CONCLUSION AND FUTURE WORK

From our experimentations, we can conclude that there is a correlation between words in an initial message and the average length of a conversation. After training 4 different machine learning models, we were able to use our logistic regression model to correctly predict the average conversation length 59.24% of the time. Machine learning helped us analyze vast amounts of profile data to identify which features contributed most to success. This can help users improve their chances of landing a date and also help us get a deeper understanding of social interaction and attraction in a digital context. Our research stands at the intersection of technology, psychology, and sociology, investigating the digital behaviors that drive human connections. Our findings suggest that users of digital dating platforms can enhance their success by tailoring their initial messages; specifically, incorporating personalized greetings and questions that invite dialogue may increase conversation lengths. Future research could explore the profile imagery and user demographics on conversation lengths in digital dating.