“How you doin’?”: Using logistic regression for character identification on Friends

Cecilia Knaub

May 22, 2025

1 Introduction

The television show Friends is one of the most beloved and culturally significant sitcoms in recent memory. The show followed six young adults living in New York City through the trials and tribulations of finding love, establishing careers, and building community. It ran for ten seasons from 1994 to 2004, broadcasting 236 episodes and endearing the main cast to millions of viewers. With such popularity and staying power, it is worth exploring how their unique character identities are communicated through their dialogue. Using a logistic regression model to review every line of dialogue in the series, this paper seeks to determine if perceptible differences in the main characters emerge through the words they speak.

Applying a language model to this question helps us understand the character traits, relational dynamics, and thematic elements that become cultural reference points in popular media. These tropes can provide insight about what constitutes compelling narrative media in contemporary television. Logistic regressions’ ability to reveal interpretable features makes this a valuable application from a cultural analytics perspective.

The dataset, created by Jinho D. Choi and the Emory NLP team, contains all spoken dialogue from the series, attributing each utterance to a speaker. It contains 67,373 utterances across 700 speakers. The six main characters, Chandler Bing, Joey Tribbiani, Monica Geller, Phoebe Buffay, Rachel Green, and Ross Geller, comprise only 0.85% of the speakers, yet account for 76.16% of the utterances. Given the speaker class imbalance, and this inquiry’s interest in the six main characters alone, the first step in the test processing task was to filter out all other speaker instances. Variance still exists within the six main characters, which may be a factor performing the classification task. The dataset used in the logistic regression ultimately contains 51,312 utterances from six speakers. The tsv file is loaded into Python and stored as a dataframe using the Pandas library.

The next step is to pre-process the utterance text data. The script first performs entity recognition to identify and filter out named entities using Spacy’s default parts-of-speech tagging. The text is then lemmatized, also using Spacy

Character	Utterances
Rachel Green	9331
Ross Geller	9161
Chandler Bing	8568
Monica Geller	8498
Joey Tribbiani	8215
Phoebe Buffay	7539

Table 1: The total number of utterances for each character

defaults, case-folded, and tokenized. With tokenization complete, the program filters out English language stopwords loaded from the Spacy library, and removes any highly domain-specific tokens.

The program deploys the scikit learn module for multinomial classification and regression. The text is randomly split into a training set of 41,049 utterances, a development set of 5,131, and test set of 5,132. An initial comparison reveals the development baseline probability of 18% and testing baseline probability of 17% slightly exceeds the random prior probability of 16%, indicating that the class imbalance may not impact the classification task. A hyperparameter search determines a regularization coefficient of 2.0 yields the best results on the development set.

2 Data

The model achieved 28.73% accuracy on the development set and 28.90% on the test set. With about a 55% improvement on the baseline, this suggests the model detects some patterns to aid in predicting the speaker.

Metric	Development	Test
Baseline	0.18	0.17
LR	0.2873	0.2890

Table 2: The baseline and model accuracy

Additional performance metrics support the assertion that the model detects some patterns while ultimately struggling to distinguish one character from another. The micro classification precision of 28.90% suggests the model moderately predicts which character is speaking a given utterance. Comparison of the micro and macro metrics indicates that this is done without much variance across the six characters. The micro recall metrics, equal to that of the micro precision, indicate that the model struggles to identify a character's dialogue.

A confusion matrix helps visualize the model's performance and the overlap in predictions. When the model assigns the correct speaker, it predicts similar numbers for one or more others. This is the case for Joey, Monica, Rachel,

Metric	Micro	Macro
Precision	0.2890	0.2919
Recall	0.2890	0.2865
F1	0.2890	0.2847

Table 3: Precision and Recall metrics

and Ross. For Chandler and Phoebe, the model is more likely to assign their dialogue to other characters, revealing the accuracies in the model's predictive ability.

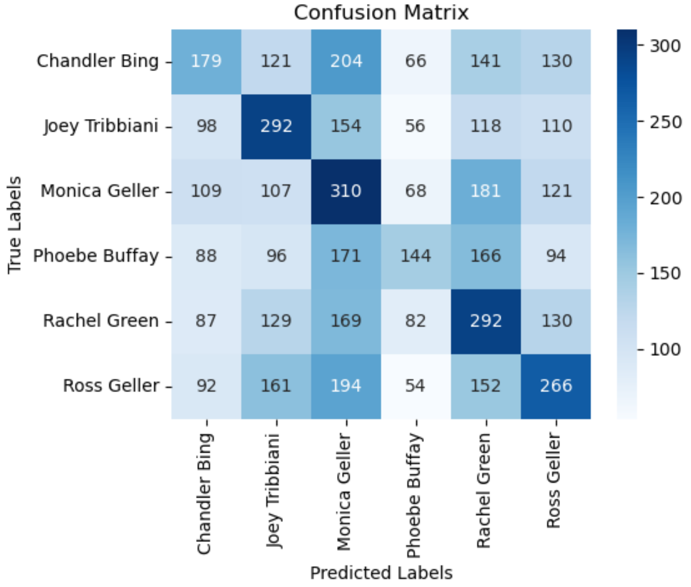

Figure 1: A confusion matrix showing showing the true and predicted class by the model

An analysis of the learned coefficients reveals the most positively and negatively associated words with each main character. Utterances may be interpreted as personality traits, motivations, notable plot points, themes, or other stereotypical cues describing the speaker class. One of the more common positive associations provide insight on what the characters do for work: *angle* (4.4355) and *oscar* (3.9221) may relate to Joey's acting career, *cater* (3.5706) relates to Monica as a chef, *gala* (3.6257) and *vogue* (3.5858) are relevant to Rachel's career in fashion. These indicate that one of the distinguishing characteristics of the main cast is their careers and that those details factor into events in the series.

See Table 4 on page 5 for the all associated tokens.

3 Conclusion

The modest performance in character identification suggests overlap in dialogue patterns among the main characters. Consistent accuracy, perception, and recall indicate that the dialogue alone is not unique enough to identify the speaker at a high probability. A review of the learned coefficients reveals that careers are one of the few differentiators among the characters. A few factors could explain these results and inspire additional research into understanding popular characterizations in contemporary television. First, the model's consistent, low accuracy in predicting a speaker may suggest that the distinctiveness of beloved characters often emerges from a complex interplay of factors beyond just the words they speak. Analyzing the dialogue alone does not consider the visual or audio elements of the medium, among an assortment of possible factors, both of which may contribute to compelling characterizations. Additionally, Friends relies on six characters to develop the emotional stakes and drive the plot action. The ensemble nature of the show results in shared conversation dynamics. These findings may reveal an intentional storytelling approach intended to create a cohesive group identity that audiences find relatable and engaging. Furthermore, these findings suggest that ensemble sitcoms may depend more on situational comedy and interpersonal relationships than on distinctive characterizations. Future research should explore approaches that integrate different linguistic analysis, or a multimodal approach considering visual and audio features to better understand how ensemble television characters achieve distinctiveness while creating a convincing only realistic group of friends.

Chandler Bing	**Joey Tribbiani**	**Monica Geller**	**Phoebe Buffay**	**Rachel Green**	**Ross Geller**
clench (4.4289)	aaron (5.5464)	rache (4.2662)	opponent (6.3941)	demand (4.5182)	los (4.5545)
mascara (4.2791)	came (4.9100)	puke (4.2175)	abott (4.6846)	dirt (4.2804)	arrive (4.4847)
nina (4.1851)	angle (4.4355)	mockolate (4.1181)	patient (4.6642)	digital (4.1713)	ouch (3.8915)
deposit (4.0480)	oscar (3.9221)	shock (4.0708)	philange (4.4997)	yeti (4.0965)	hans (3.8215)
heldi (4.0136)	eva (3.9160)	slick (4.0642)	minsk (4.4917)	tiney (4.0823)	evidently (3.7568)
jelly (3.8767)	neurologist (3.8044)	java (3.9233)	sergei (4.2037)	cart (4.0684)	execute (3.6520)
israeli (3.8700)	comin (3.6978)	humidity (3.8680)	pesos (4.0472)	zelner (3.8469)	bitty (3.4574)
iffy (3.7748)	dork (3.6853)	seventeen (3.7592)	consuela (3.8378)	gala (3.6257)	rage (3.2336)
saying (3.7364)	confusion (3.4591)	plain (3.6719)	suicide (3.7821)	contract (3.5921)	sfunny (3.2232)
refundable (3.6238)	wayne (3.4501)	cater (3.5706)	stab (3.7545)	vogue (3.5858)	mill (3.2049)
bar (-2.8894)	honey (-3.4906)	grab (-3.5197)	devastate (-4.2414)	richard (-3.5573)	similar (-3.4997)
honor (-2.8638)	american (-3.3554)	medium (-3.3183)	idiot (-2.8676)	erica (-3.0118)	invitation (-3.0344)
barn (-2.4219)	sec (-3.3521)	chick (-2.9543)	amy (-2.8345)	cole (-3.0058)	storm (-3.0122)
pervert (-2.4161)	dyou (-3.2914)	opponent (-2.8764)	phoebs (-2.8198)	choose (-2.8795)	tight (-3.0114)
barry (-2.4078)	plane (-3.0932)	ignore (-2.8706)	exchange (-2.7829)	favor (-2.8591)	francis (-2.9415)
faith (-2.3877)	birdie (-2.9941)	choice (-2.7938)	lifetime (-2.6752)	noisy (-2.8152)	kitty (-2.9022)
mona (-2.3689)	bud (-2.8901)	buffay (-2.7321)	camera (-2.6646)	bill (-2.7298)	bite (-2.7847)
magnet (-2.3264)	bike (-2.6942)	client (-2.6900)	letter (-2.6265)	kathy (-2.5780)	bright (-2.7806)
vulnerable (-2.2800)	yay (-2.6220)	dude (-2.6173)	pheebs (-2.6139)	furniture (-2.5239)	bonus (-2.7556)
purse (-2.2321)	opener (-2.5322)	becker (-2.5922)	jellyfish (-2.5637)	gold (-2.5007)	explanation (-2.6885)

Table 4: Tokens associated with each character