

A photograph of two individuals from the waist up, standing side-by-side against a solid blue background. On the left, a person with dark skin is wearing a bright green V-neck sweater over a pink collared shirt and orange tie, paired with blue shorts. On the right, another person with dark skin is wearing a bright yellow V-neck sweater over a green t-shirt, paired with a magenta skirt.

SAHAJ CHAWLA - DECEMBER 2020

SUBREDDIT CLASSIFICATION

DSI-18 Project 3

**WHAT ARE THE INDICATIVE KEYWORDS THAT HELP TO
EFFECTIVELY TARGET ADVERTISING TO A NICHE USER GROUP?**

WHICH SUBREDDITS WILL I LOOK INTO AND WHY?

About Community

...

Welcome to Tales From Call Centers (TFCC), a place where we share tales from the trenches of the call center world!

This includes things like (but not limited to); Ridiculous caller demands Moronic and stupid things callers say Moral support after dealing with awkward and difficult callers Happy and positive calls

208k

Members

219

Online

Created 14 Aug 2012

About Community

...

Welcome to Tales From Tech Support, the subreddit where we post stories about helping someone with a tech issue. Did you try turning it off and on again?

650k

Members

675

Online

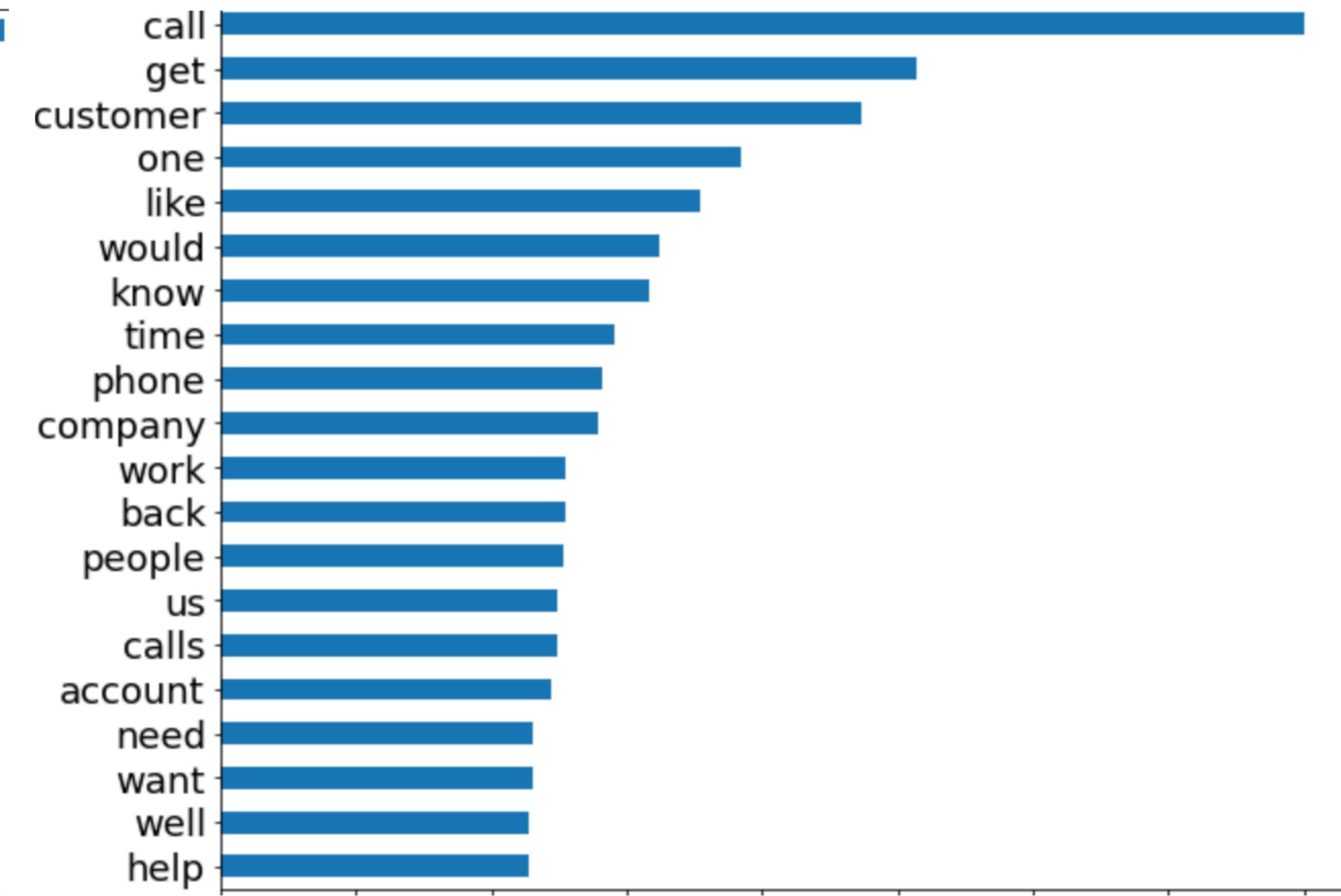
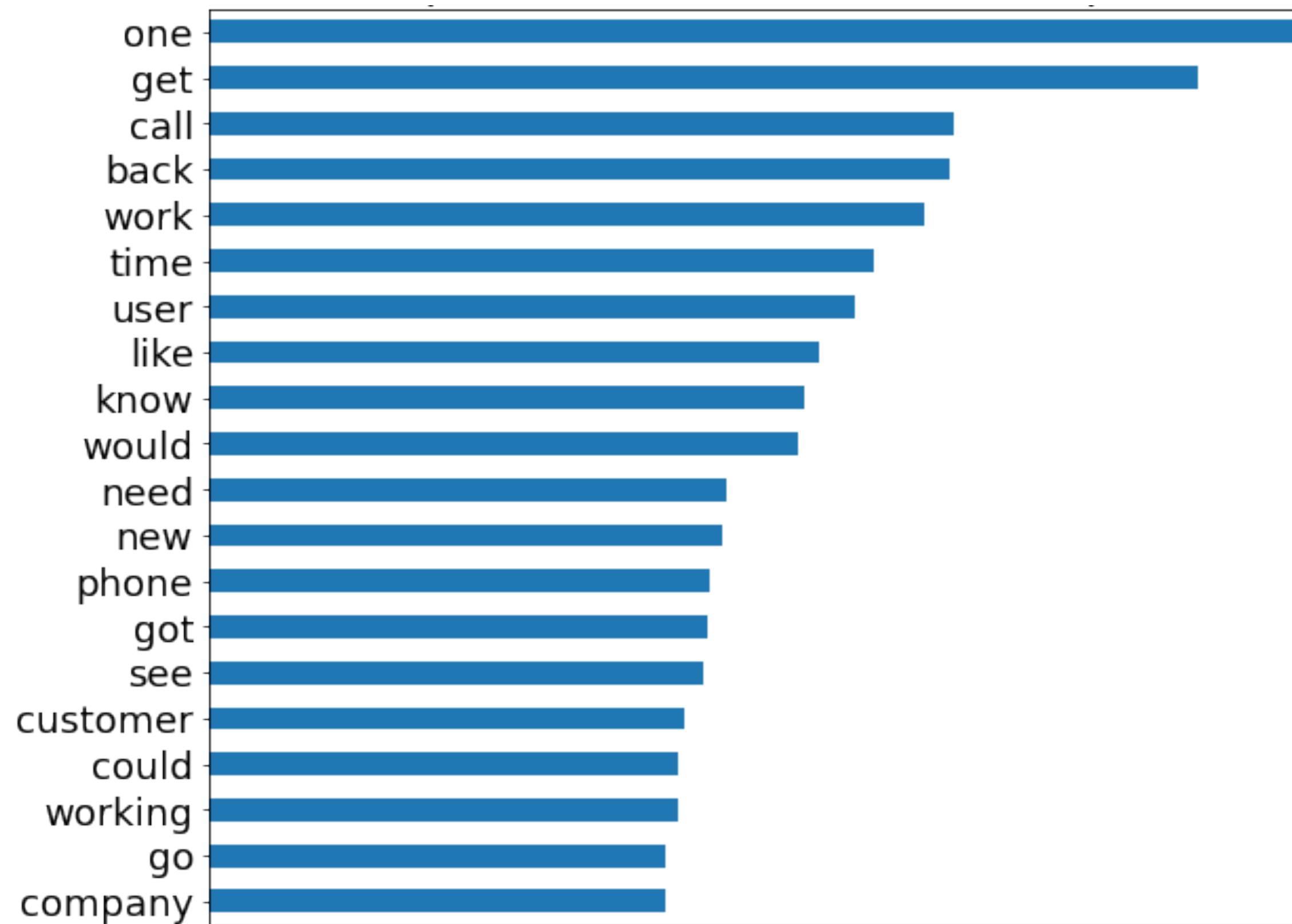
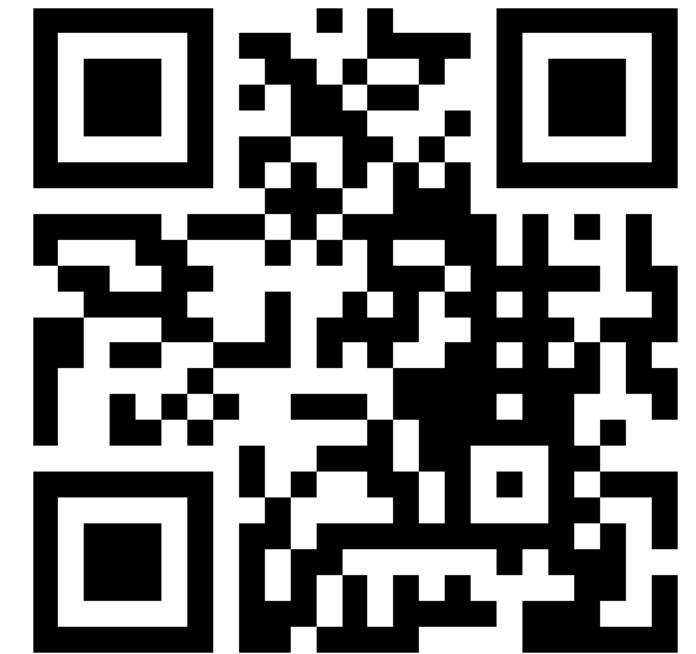
Created 12 Apr 2011

WHAT DID I DO?

- 1. Scraping the data from Reddit**
- 2. Exploratory Data Analysis**
- 3. Natural Language Processing**
- 4. (A lot of) Classification Modeling**
- 5. Scraping the data from Reddit (again)**
- 6. Model Evaluation**

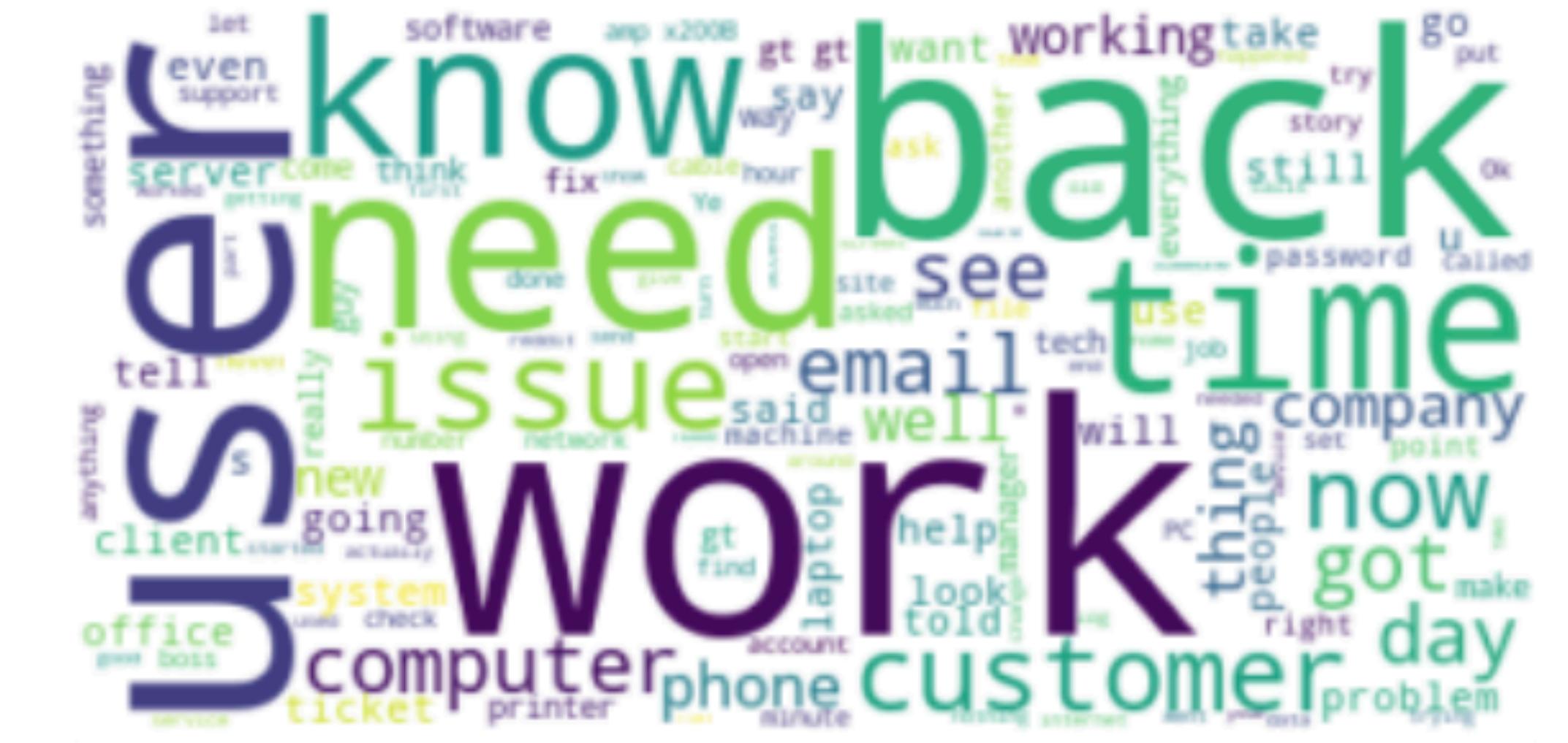
EXPLORATORY DATA ANALYSIS

FREQUENCY OF TOP WORDS IN BOTH SUBREDDITS - CAN YOU GUESS WHICH ONE BELONGS TO R/TALESFROMTECHSUPPORT?



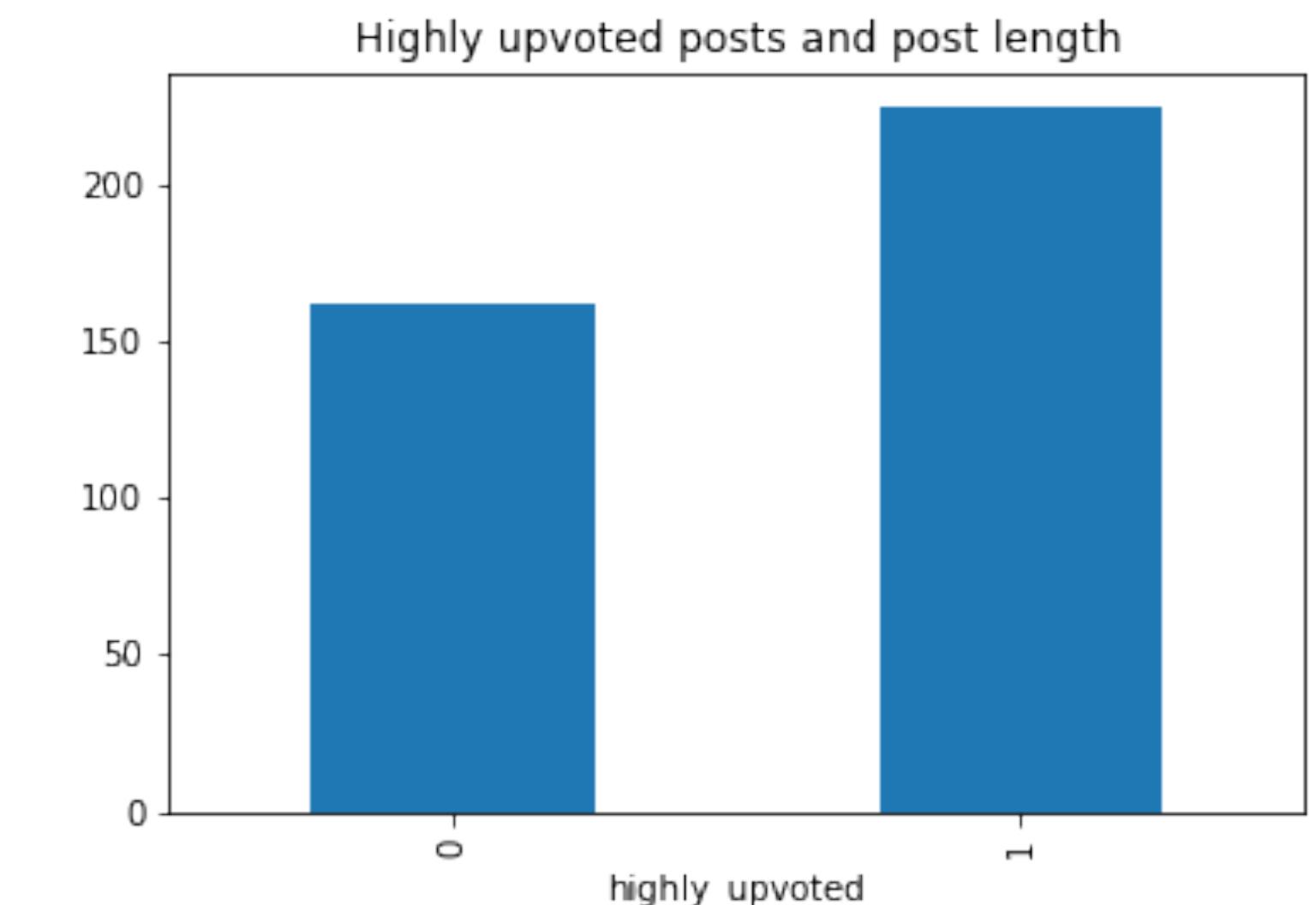
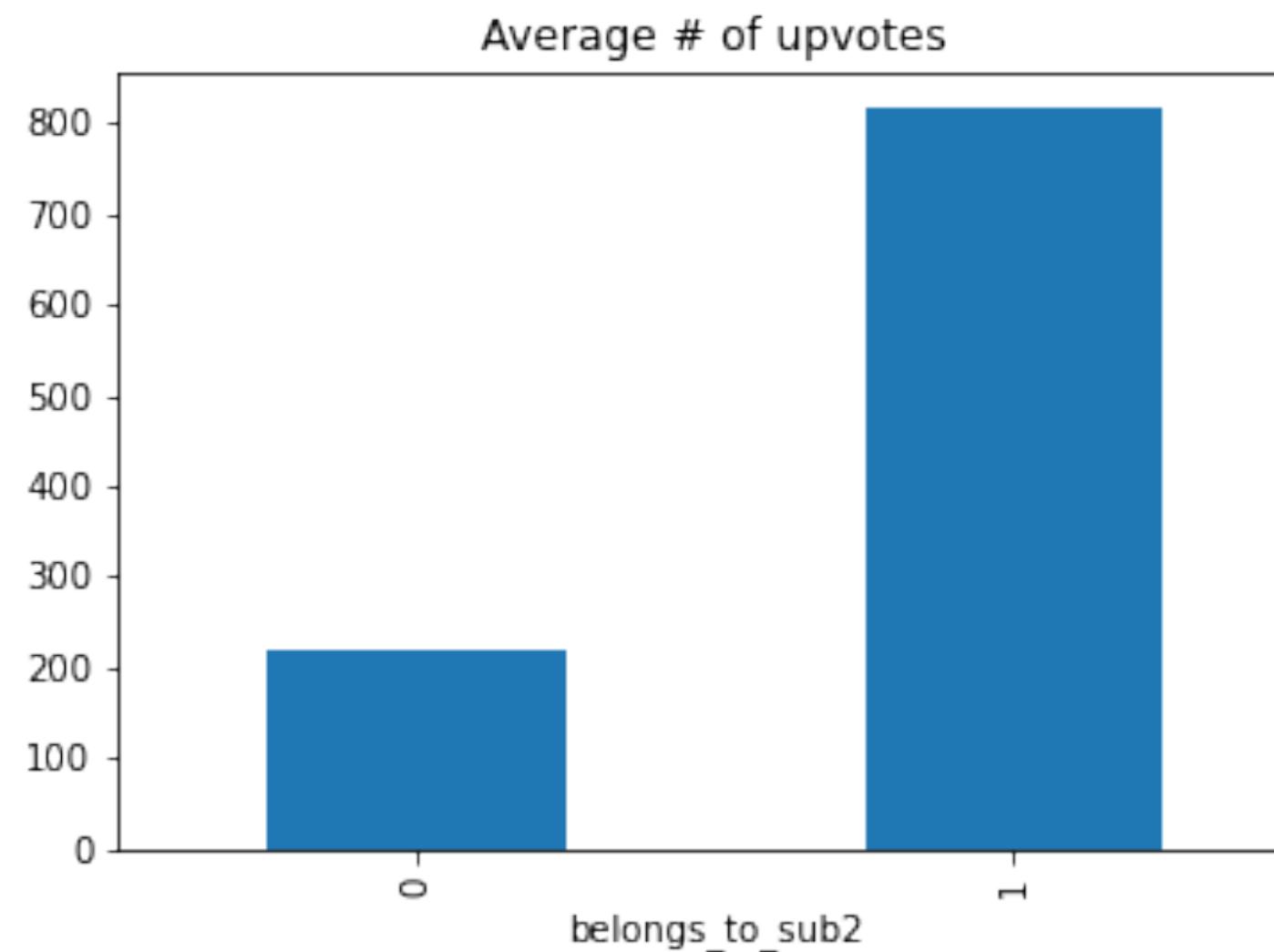
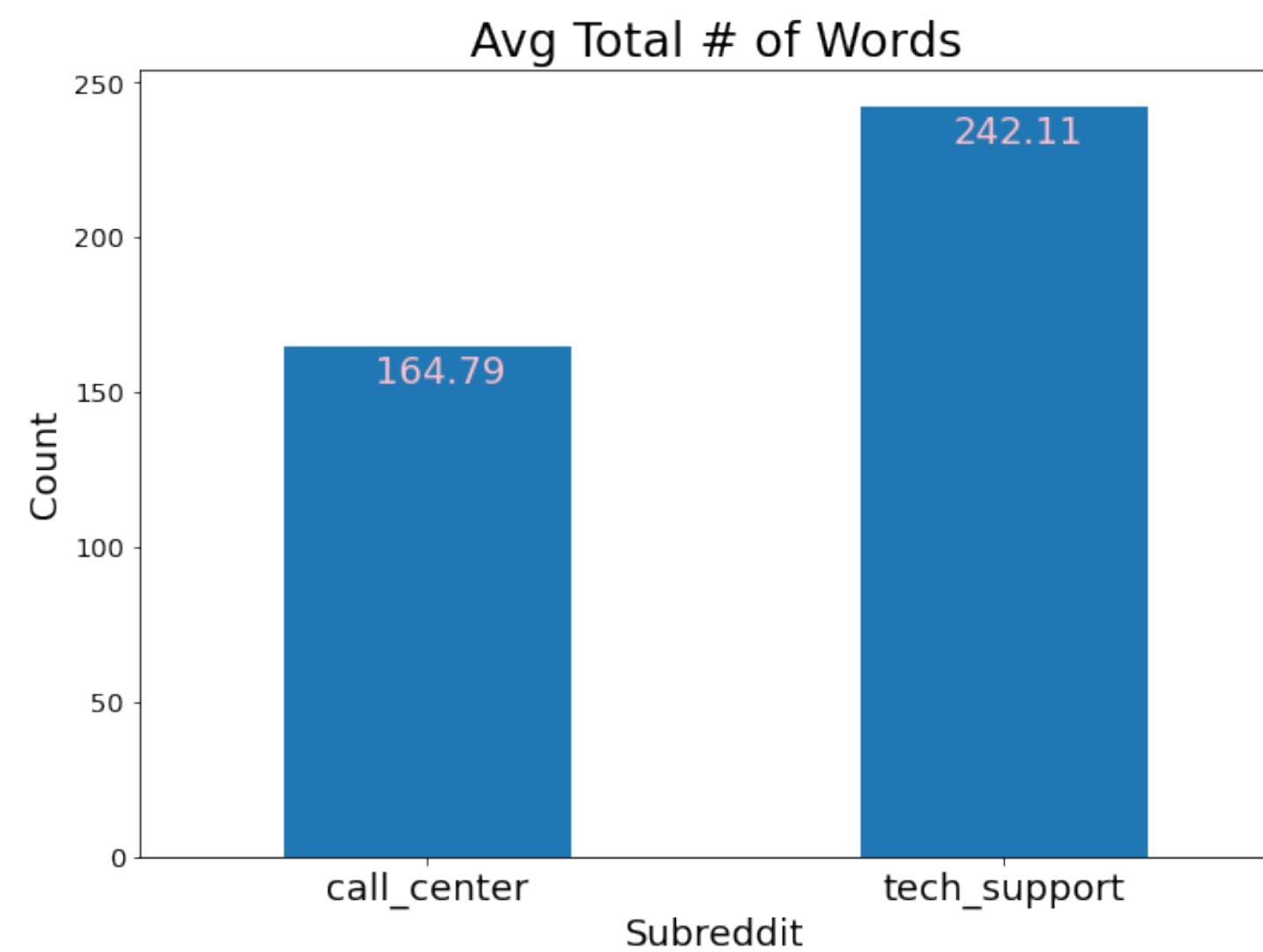
Voting options: Scan the QR code OR <https://www.menti.com/ehm3if9iyb> OR go to menti.com and use the code 77 36 08 0

CONTENT OF THE 2 SUBREDDITS SEEMS TO FOLLOW A SIMILAR THEME, BUT THERE ARE CLEAR DIFFERENCES IN NUANCES



EXPLORING THE RELATIONSHIP BETWEEN POST LENGTH AND UPVOTES

- Post length and average number of upvotes is different across the 2 subs
- Highly upvoted (more than 100) posts tend to have a greater effort put into them (higher word count)



BONUS: LATENT DIRICHLET ALLOCATION FOR TOPIC MODELLING VISUALISATION

Topic 1

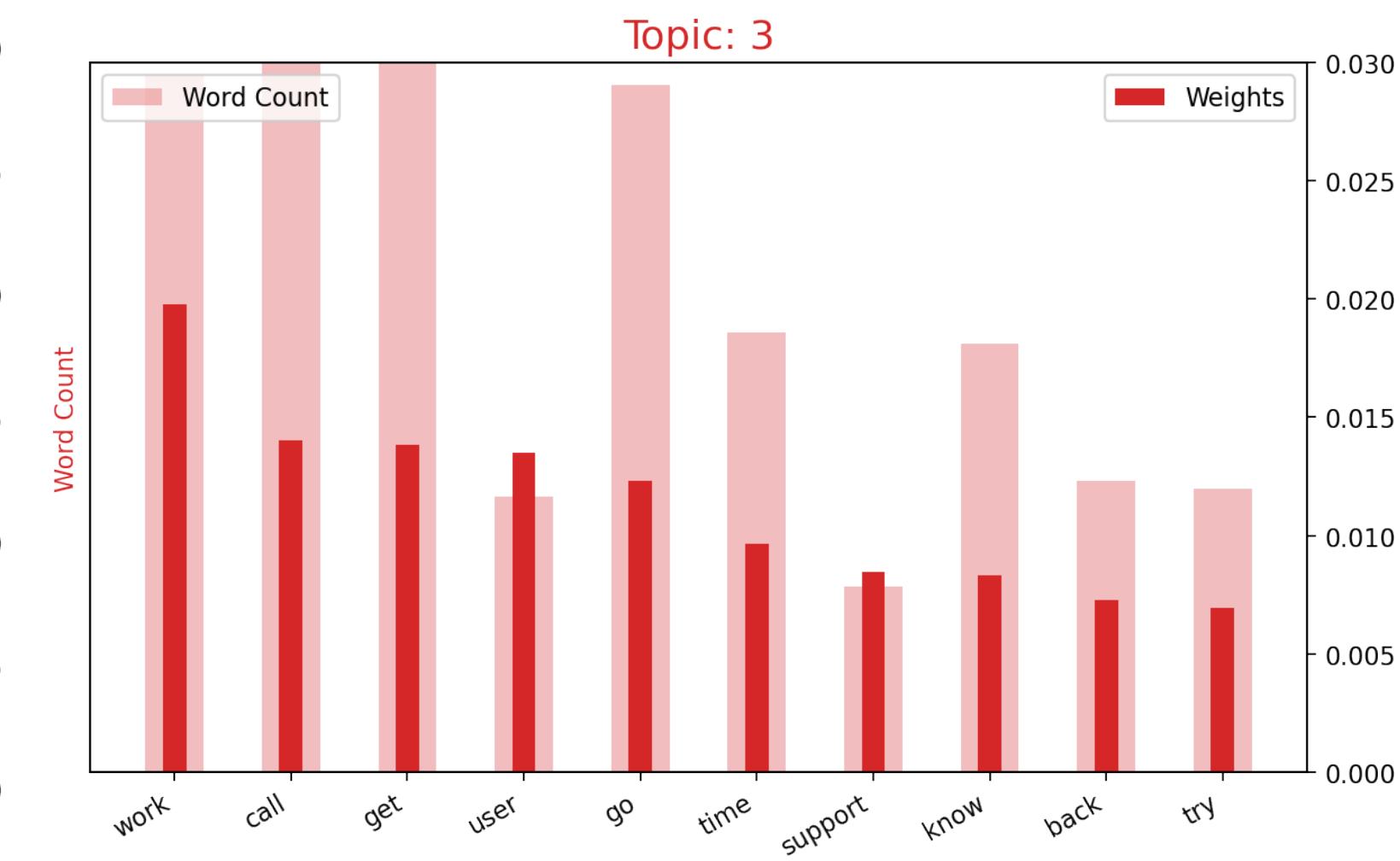
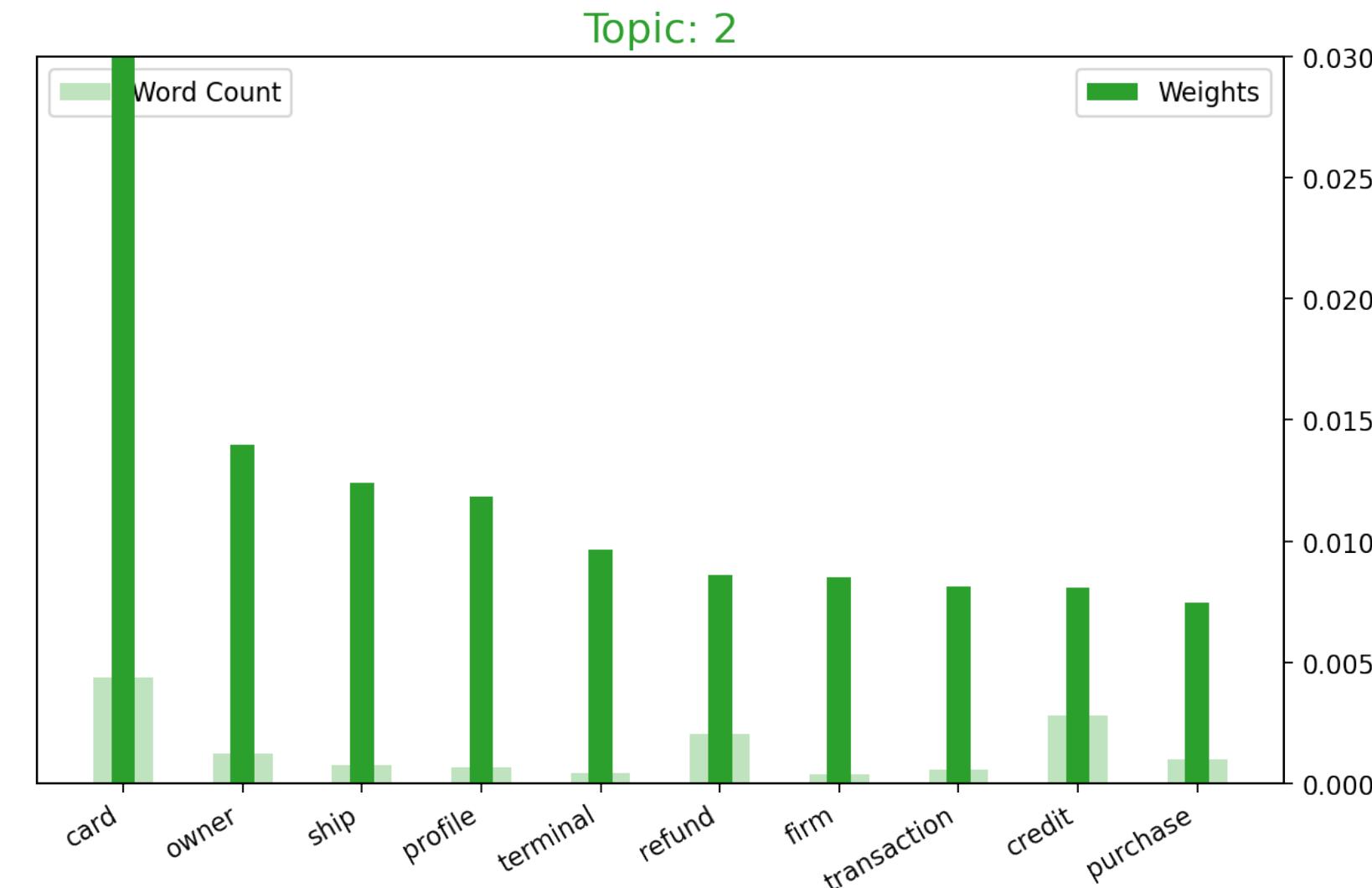
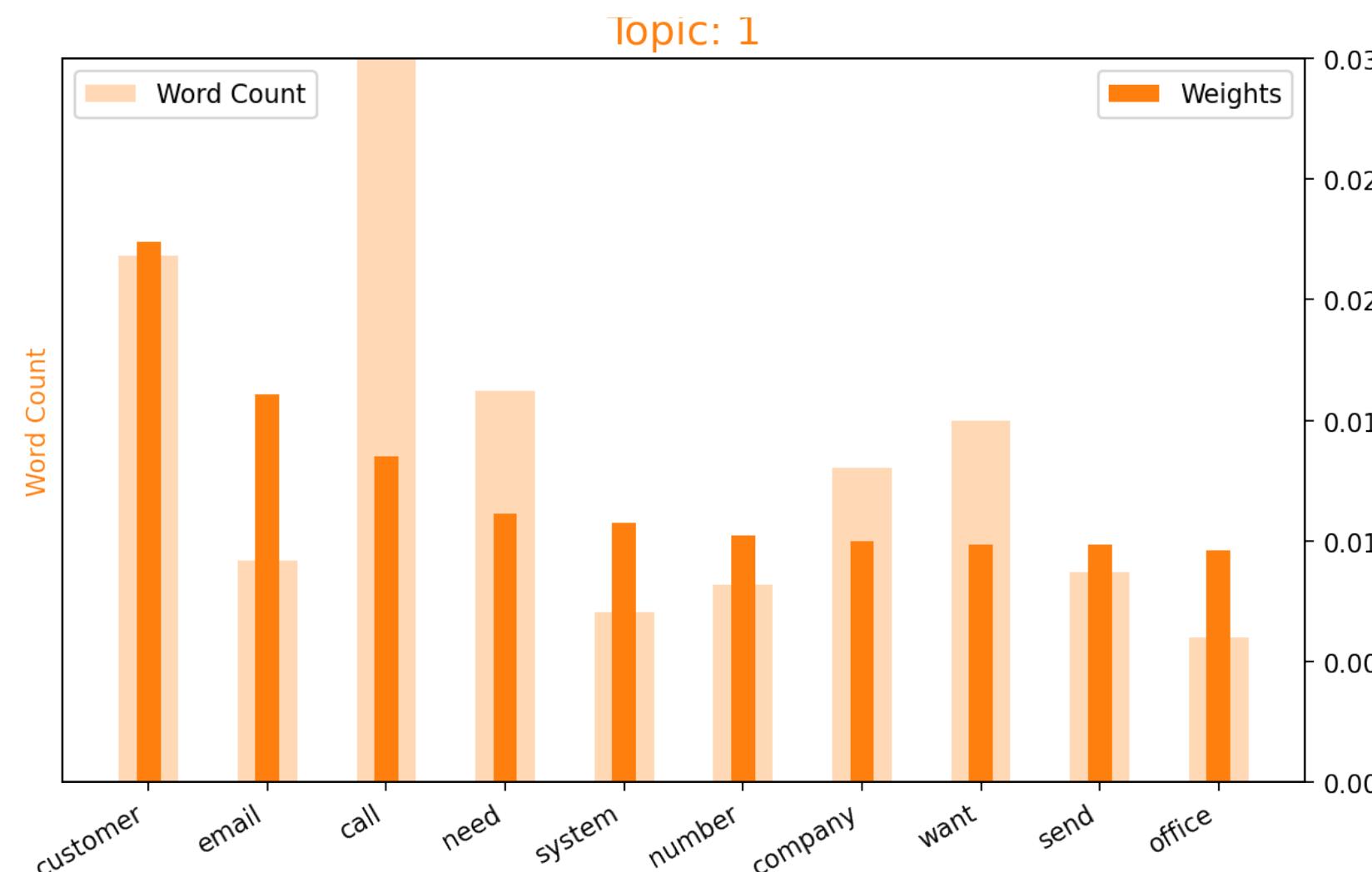
customer
call need
email send office
system company
number

Topic 2

firm transaction
ship credit
owner
refund card
profile purchase
terminal

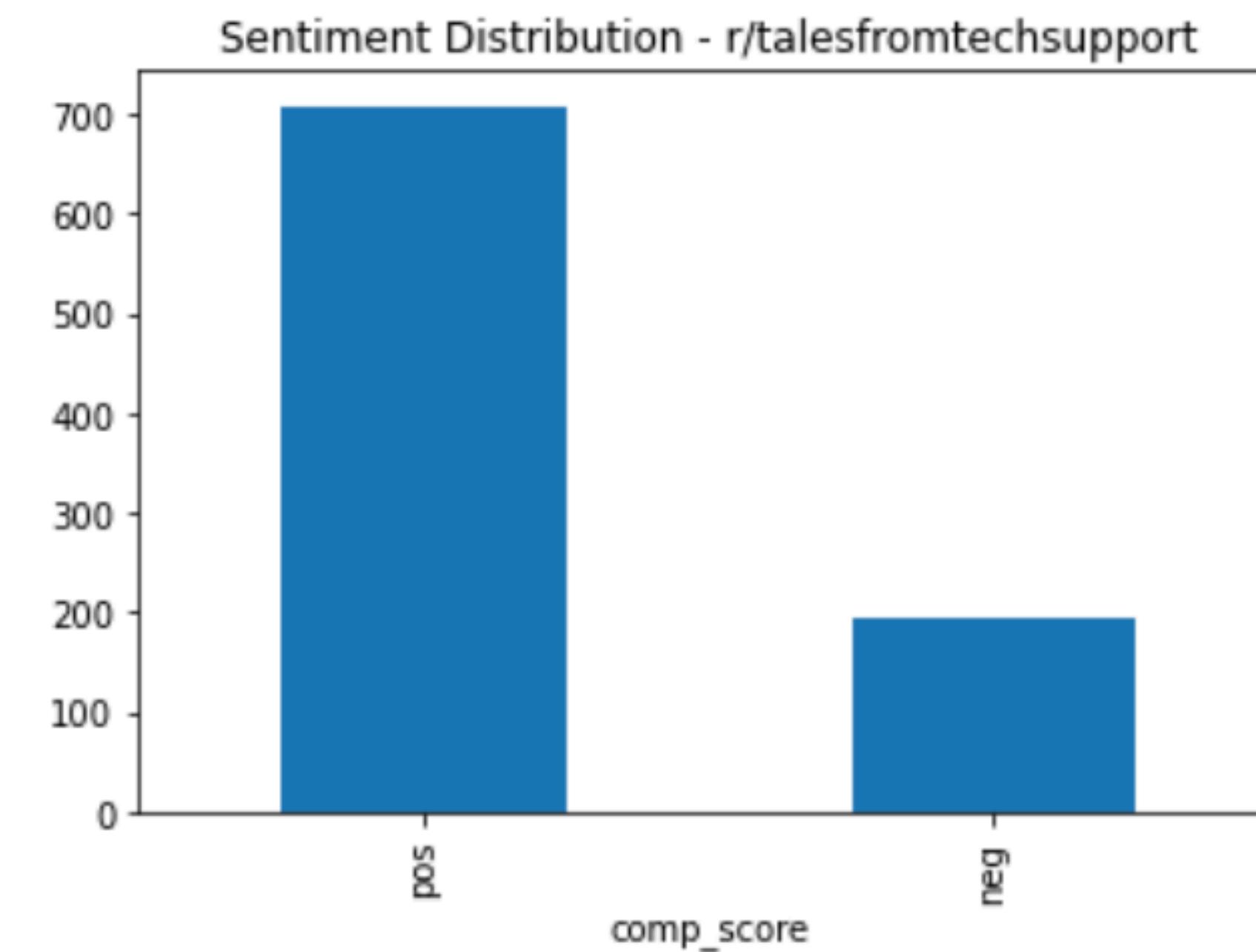
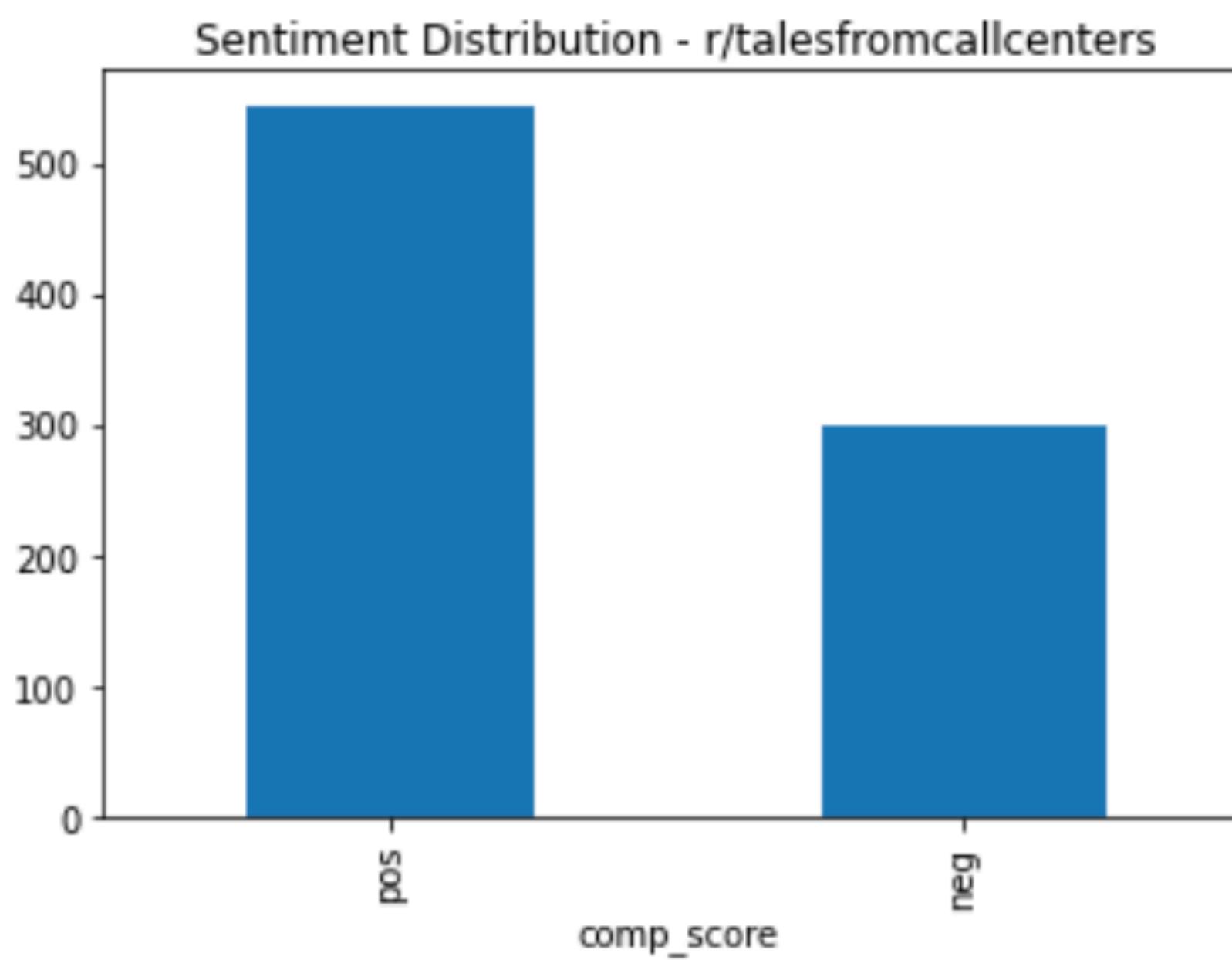
Topic 3

go support
call user
know time get
work try back



DOUBLE BONUS: SENTIMENT ANALYSIS WITH VADER LIBRARY

- Not much difference in sentiment between the two
- No discernible relationship between sentiment and upvotes



PRE-PROCESSING AND DATA MODELLING

NATURAL LANGUAGE PROCESSING

- Dropped duplicate posts
- Joined title and post to create one string
- Converted all to lowercase
- Removed punctuation
- Lemmatized words
- Removed stopwords

	title	post	upvotes	gilded	belongs_to_sub2	title_x_post	token	joined_text	highly_upvoted	word_count	char_count	cleaned_text
0	Just need to vent...	I get that tempers are shorter these days, but...	230	0	0	Just need to vent... I get that tempers are sh...	['need', 'vent', 'get', 'tempers', 'shorter', ...]	Just need to vent... I get that tempers are sh...	1	48	579	need vent get tempers shorter days hard time l...
1	Reverse call center post	On mobile so I hope I do this right.\n\n ha...	39	0	0	Reverse call center post On mobile so I hope I...	['reverse', 'call', 'center', 'post', 'mobile'...]	Reverse call center post On mobile so I hope I...	0	84	1017	reverse call center post mobile hope right cal...
2	"So you're willing to lose a customer for \$3 d...	I work for a car rental company as a specialis...	763	0	0	"So you're willing to lose a customer for \$3 d...	['willing', 'lose', 'customer', 'dollars', 'wo...	"So you're willing to lose a customer for \$3 d...	1	585	6757	willing lose customer dollars work car rental ...
3	Free Talk Friday - Nov 27	Welcome to Free Talk Friday! We are suspending...	0	0	0	Free Talk Friday - Nov 27 Welcome to Free Talk...	['free', 'talk', 'friday', 'nov', 'welcome', '...]	Free Talk Friday - Nov 27 Welcome to Free Talk...	0	38	368	free talk friday nov welcome free talk friday ...
4	Accidentally Exposed a Family Fraud	I work for a small local ISP. One of the thin...	958	0	0	Accidentally Exposed a Family Fraud I work for...	['accidentally', 'exposed', 'family', 'fraud',...]	Accidentally Exposed a Family Fraud I work for...	1	366	3608	accidentally exposed family fraud work small l...

MODELLING PIPELINE AND RESULTS OVERVIEW - VALIDATION SET

- Vector transforming the data with CountVectorizer/TF-IDF
- Use GridsearchCV to optimise over a range of models and hyperparameters

Best Models	Accuracy Score	Estimator	Hyperparameters
LogReg	92.8%	TF-IDF	<code>tf_max_features': 10000, 'tf_ngram_range': (1, 2)</code>
Naive Bayes	93.4%	CVEC	<code>cvec_max_df': 0.9, 'cvec_max_features': 10000, 'cvec_min_df': 3, ' cvec_ngram_range': (1, 2)</code>
Random Forest	93.1%	TF-IDF	<code>tf_max_features': 10000, 'tf_ngram_range': (1, 2)</code>
Extra Trees	93.4%	TF-IDF	<code>tf_max_features': 10000, 'tf_ngram_range': (1, 2)</code>
SVM	93.7%	TF-IDF	<code>tf_max_features': 10000, 'tf_ngram_range': (1, 1)</code>

Baseline accuracy: 51%

MODEL EVALUATION WITH ‘TRUE’ TEST DATA

USING A FRESH SCRAPE AND TREATING AS ‘TRUE’ UNSEEN DATA

Best model with fresh scrape - Random Forest with TF-IDF Vectorizer

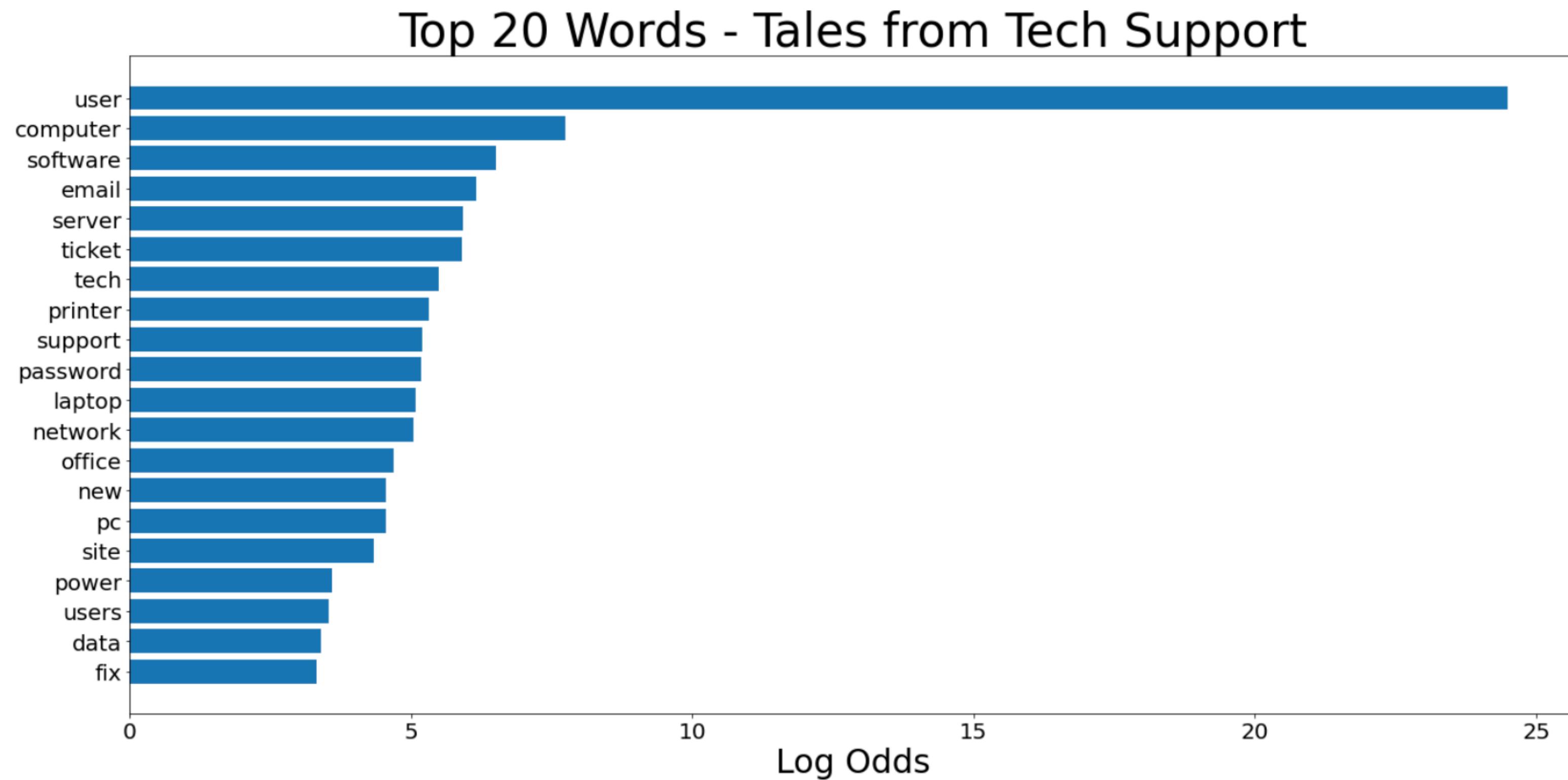
	Predicted Negative	Predicted Positive
Actual Negative	398	2
Actual Positive	0	401

- This is close to 99% accuracy, the increase is due to not enough time between the 2 scrapes.
- Ideal scenario will be to scrape more data and 2 ‘dips’ over a longer gap for a better real-life accuracy

**BUT DOES THE ‘BEST’ PERFORMING MODEL
ANSWER THE BUSINESS QUESTION?**

AN INTERPRETABLE MODEL ANSWERS THE BUSINESS QUESTION BETTER

- With the highest accuracy score on 'true' unseen data, Random Forest is theoretically the best model to answer the data science challenge.
- However, Logistic Regression is a better way to answer the business problem, as the model is interpretable and provides relevant keywords for better classification.



APPENDIX

LIMITATIONS AND SCOPE FOR IMPROVEMENT

- Longer time period between scrape dips to avoid duplicity
- Better data cleaning to remove more stop words
- Better hyper parameter tuning
- More grid searching to optimise hyper parameters

NAIVE BAYES COEFFICIENTS

Top 20 Words - Tales from tech support

