DRG Final Report

Team: Algorithmic Sensemaking

Nayan, Xiaobing

Due December 8th, 2022

Algorithmic Sensemaking

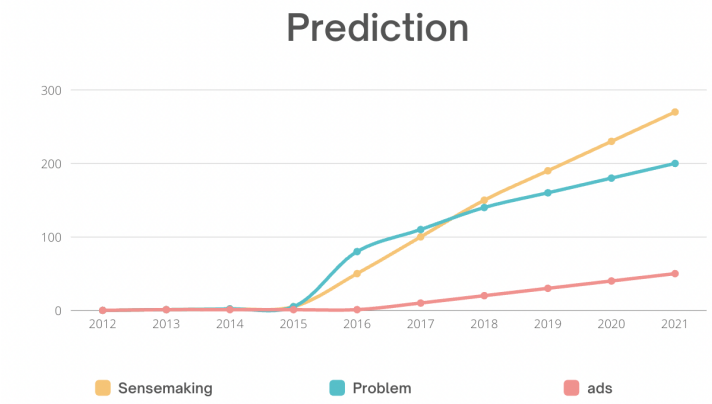
Overview

This quarter you worked with some of the labeled data. Use this worksheet to reflect on progress, next steps, and goals for next quarter. Follow the prompts and include visualizations when specified! Thank you so much for all your hard work – this is a difficult problem that has not been very successful yet. Thank you for your continued effort and good attitudes. Achieving an accurate classifier will be a huge step for both this work, and for your CVs!

Algorithmic Literacy prediction

1. What do you predict will be the case of how algorithmic literacy has changed over time? In other words, how will the relative frequencies of “sensemaking”, “problem”, and “ads” change from 2015(ish) to 2021 and beyond? Justify your prediction.

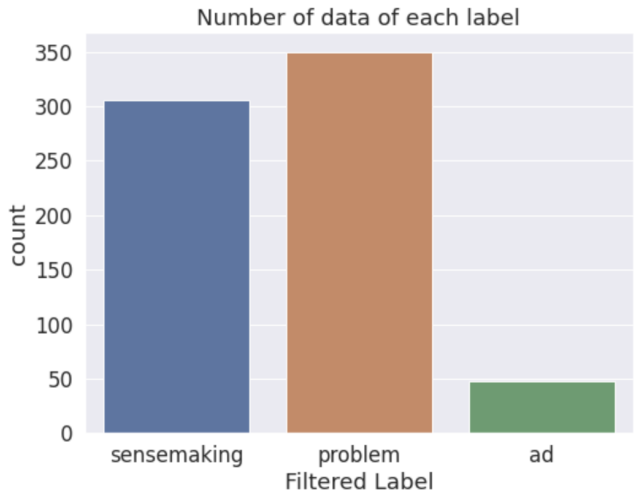
From our prediction, we believe that the number of sensemaking and ads will have a higher growth rate compared with the number of problem posts from 2015 to 2021. For the future,



the trend will stay the same.

Raw Numbers

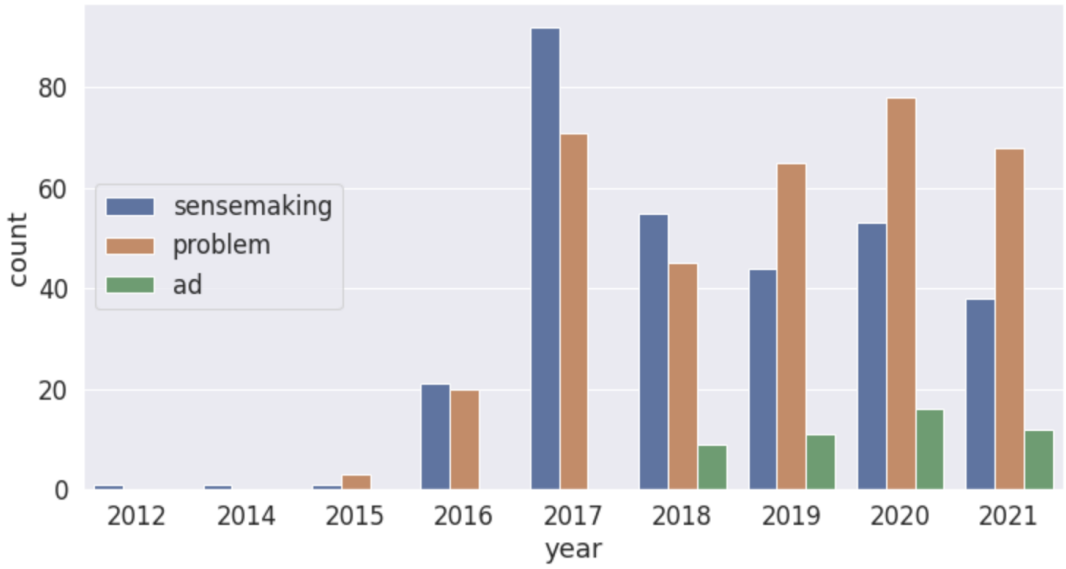
2. How much data do we have? What is the distribution of the labels (sensemaking, problem, ad) as well as the distribution of dates and labels over time? Show this in bar charts, as many as you think are relevant to answer this question.



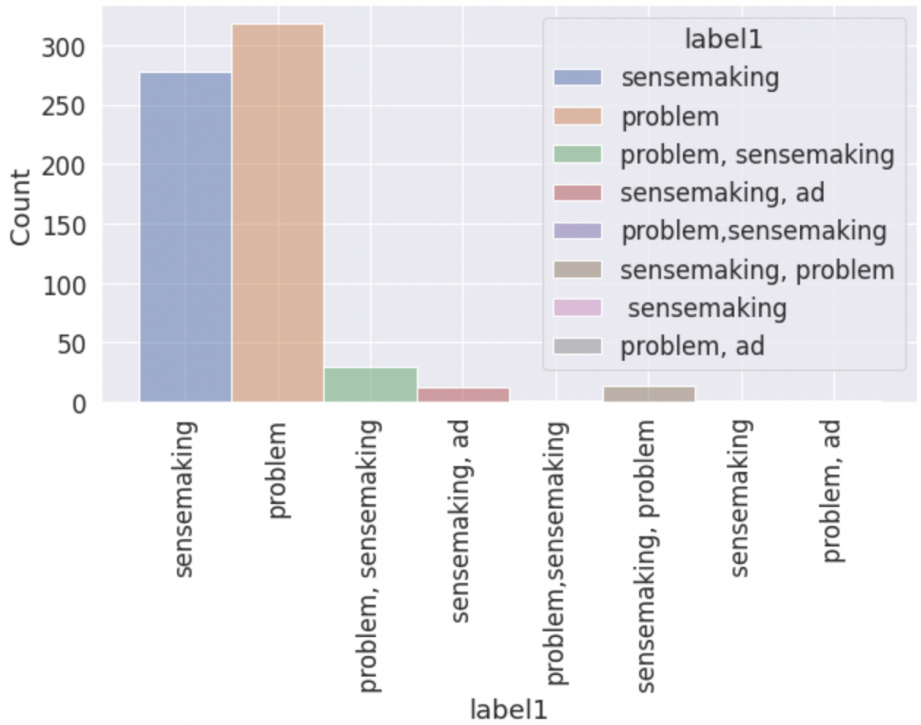

```
df['label_filter'].value_counts()
```

```
problem        350
sensemaking    306
ad              48
Name: label_filter, dtype: int64
```

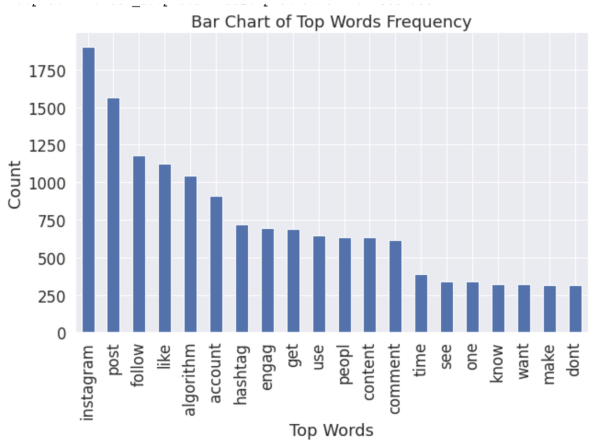
After data cleaning and preprocessing, the total number of valid data is 704. The number of data for sensemaking is 306, the problem is 350, and the ad is 48.



We have more labels of sensemaking than problems compared with problem on 2016, 2017, and 2018. More problems show up in 2019, 2022, and 2021. The reason of causing the difference between the graph and our predictions might relate with label preprocessing. The original labels are messy, and we want to make it consistent in 3 categories, sensemaking, problem, and ad.

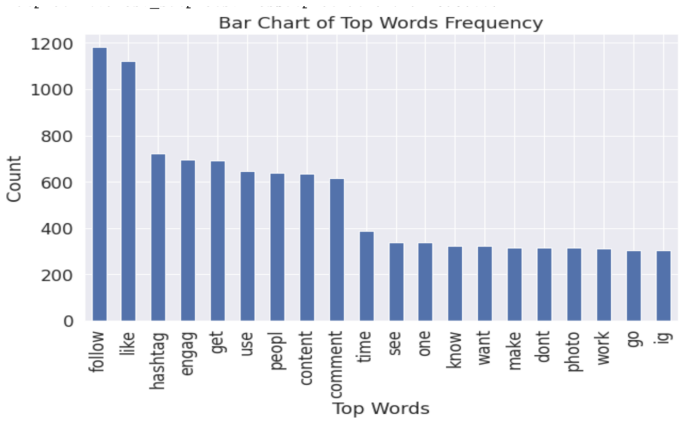


3. What are the top words for each label? Justify how you chose to show this. What about after we remove common words like “post”, “account”, or “algorithm”?



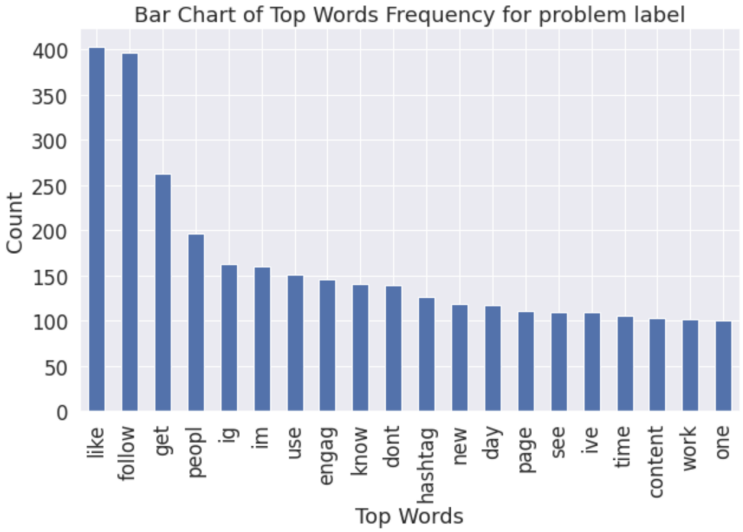
The plot above is before removing common words like ‘post’, ‘account’, ‘algorithm’, and ‘instagram’.

	Post	count
0	follow	1181
1	like	1122
2	hashtag	721
3	engag	694
4	get	690



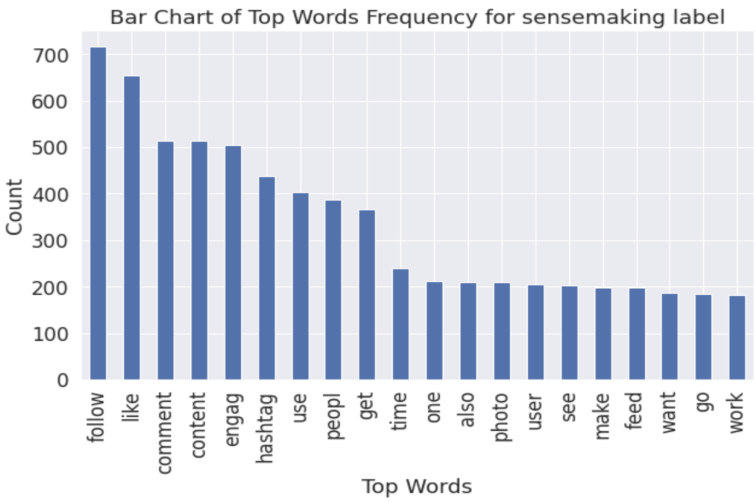
The plot above is after removing common words for all labels.

	Post	count
0	like	403
1	follow	396
2	get	263
3	peopl	197
4	ig	163



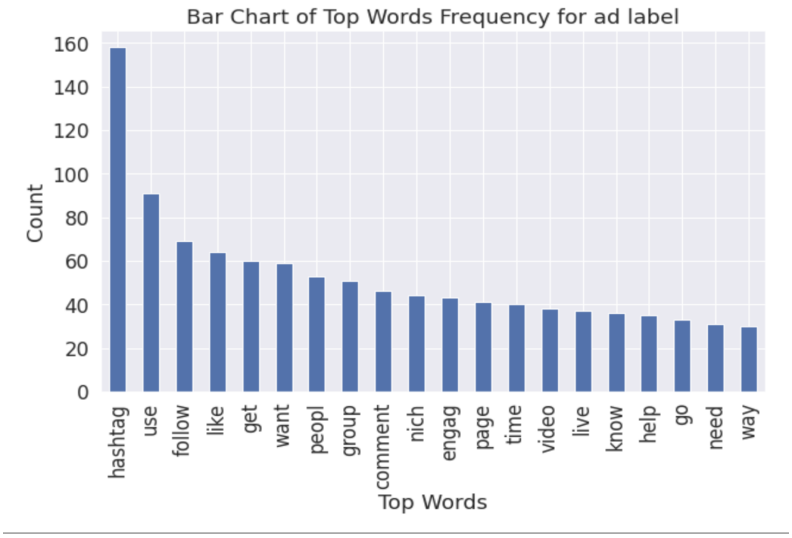
Top words of the post with problem label are ‘like’, ‘follow’, ‘get’, ‘people’

	Post	count
0	follow	716
1	like	655
2	comment	514
3	content	513
4	engag	505



The top words of the post with the sensemaking label are ‘follow’, ‘like’, ‘content’, ‘engage’, ‘hashtag’.

	Post	count
0	hashtag	158
1	use	91
2	follow	69
3	like	64
4	get	60



Top words of the post with the ad label are ‘hashtag’, ‘use’, ‘follow’, ‘like’, and ‘get’.

4. Looking at examples online for Naive Bayes or other algorithms, how much data is typically required for a classifier to work? For example, product reviews or credit evaluation data.

2000 data points

Classification Models

5. Explain what a language classification model does in simple terms. What is a basic classification model doing, and how? What is the goal?

Language Classification models analyze textual data and classify the text into predefined classes (based on which they have been trained). These models usually require textual data to be present in a form that is interpretable for the model - e.g. the features may be words and their tf-idf matrices, or sentiments, top words etc.

6. What is the difference between Naive Bayes and a model that relies on neural networks?

1. **Naive Bayes classifier is the fastest of the bunch, but its classification accuracy will depend on the assumption that every feature is independent. The Neural Network**

requires a significant amount of computational time to train the model, depending on the number of layers and nodes in it.

2. **While Naive Bayes is easily updatable, neural networks need to be retrained after a single instance.**
3. **Neural Networks require several iterations while Naive Bayes requires just one to get the results.**

7. Report which models you tried, and what you think is going 'wrong'?

Following are the models we tried:

1. **Logistic Regression - While the training accuracy of the model was ~79.9%, which is great, we believe that given the high dimensionality of the model, it could've been due to overfitting. Currently, we're unable to test it because of the shortage of data points, but we believe that linear models would not be the best way to go forward with text classification.**
2. **Naive Bayes - Traditionally, Naive Bayes has proved to be an accurate model in classifying textual data.**
3. **BERT - BERT is a bidirectional training model which can be best used in sentence classification. Given the textual nature of our data, we believed that BERT would be the best way to go forward with the classification of the reddit posts into sensemaking, ads, and problems. However, the accuracy of the model was 55.7% which may not be good enough.**

What went "wrong":

1. **The number of data points may not be enough to develop an accurate model.**
2. **Currently, we're only using the features extracted through tf-idf for modeling, which might not be enough for classification.**
3. **We have run multiple models without entirely completing the Exploratory Analysis of the data.**

8. Report the overall accuracy for each model you tried, as well as the Confusion Matrix diagram.

Overall accuracy for each model is given below:

1. **Logistic Regression: 79.9%**
2. **Naive Bayes: 56.8%**
3. **BERT: 55.7%**

Moving Forward

9. Find a relevant paper that does language classification. Summarize what they were classifying and what they did. Reflect on what we may want to try moving forward.

A relevant paper which performs language classification is "Naive Bayes versus BERT: Jupyter notebook assignments for an introductory NLP course". It is a paper on two assignments for BSc. Data Science students at Dublin City University.

The students were required to use a dataset which classified movie reviews as "thumbs up" or "thumbs down". The first assignment was on "Sentiment polarity with Naive Bayes" to classify movie reviews as positive or negative. I believe this would be essential in adding new features (on sentiment of the tweet) into our dataset. By following the steps given in the paper, the student received a baseline accuracy of 83%.

In the second part of the assignment, the students fine-tune BERT on the same dataset using Hugging Face Transformers and PyTorch Lightning.

I believe this is an interesting paper because it provides us methods which we can use to classify our textual data and analyze our results using the models that we had previously selected (Naive Bayes and BERT).

10. Write out your next steps for next quarter for approaching this classification problem.

Next steps for the next quarter in approaching this classification problem:

1. **Adding new features like "number of swear words", "sentiment", etc. can prove to be helpful in modeling through Amazon Comprehend**
2. **Qualitatively classifying the available data (we have extracted 13,000 rows of reddit data) so that we have a larger set of data points for improving the accuracy of the model**
3. **Classifying only "problem" and "sensemaking" because the proportion of "ads" is significantly lower than that of the other two classes**
4. **Understand the variance and bias in the logistic regression model to verify if it is overfitting or not**