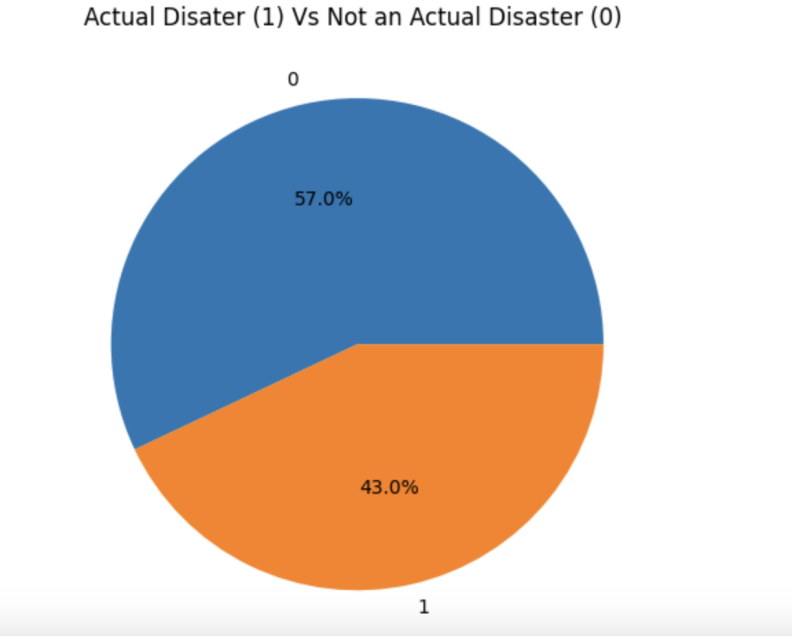
Question 1

(a) How many training and test data points are there? and (2) what percentage of the training tweets are of real disasters, and what percentage is not? Note that the meaning of each column is explained in the data description on Kaggle.

There are 7613 data points in the training data and there are 3263 data points in the test dataset. The percentage of actual disasters in the training set is 43% and not an actual disaster is 57%.



(b)

(c) I created a function to preprocess the dataset. We cleaned the text by replace any characters that are not alphabetic letters (i.e., non-alphabet characters) in the text of the current row with a space. We converted the text to lowercase to ensure uniformity in text casing. We tokenized the cleaned text by splitting it into a list of words. This tokenization is crucial for processing and analyzing text because it allows us to treat individual words or tokens as separate entities. We lemmatized the cleaned text to reduce dimensionality.We removed stopwords such as common words (e.g., "the," "and," "is") because they don't carry significant meaning.

(d) After doing trial-and-error with multiple values of M, we chose a value of 10. We thought this was a good cut-off as it got rid of unreliable features.

(e) i) The F1 score on the training data (0.872) is higher than the F1 score on the development data (0.781). In this instance, the overfitting is not severe because the F1 score on the

development data (0.781) is still reasonably high, indicating that the model is performing well on unseen data.

ii) The F1 Score on the training data is 0.8439743486936137
The F1 Score on the test data is 0.799074026681323

iii) The F1 Score on the training data is 0.8610468463423049
The F1 Score on the test data is 0.7990597572168676

iv) I think the model with L1 regularization performs the best. The F-1 score on the training data set is the lowest while the F-1 score on the development set is the same as the L2 regularization model. I think the lower F-1 score on the training data indicates that it is the least overfit.

Our logistic regression model with L2 regularization performs slightly better on the training data (F1 score of 0.861) than our logistic regression model with L1 regularization (F1 score of 0.844). When it comes to development data, our logistic regression models with L2 regularization and L1 regularization performs the same on the development data set (a F-1 score of 0.799). Compared to the logistic regression model which gave us a F-1 score of 0.872 on the training set and 0.781 on development data, the F-1 score on the training data decreased with both regularizations and the F-1 score on the development data increased. Therefore, there was some overfitting with the logistic regression model.

v) The most important words for deciding whether a tweet is about a real disaster or not are:

typhoon, Coefficient: 2.0776735592546562
outbreak, Coefficient: 1.7091813193665542
evacuated, Coefficient: 1.6251098240992359

(f) Train this classifier on the *training* set, and report its *F* 1-score on the *development* set.

The F1 Score on the training data: 0.7681596658157346
The F1 Score on the development data: 0.7320113314447593

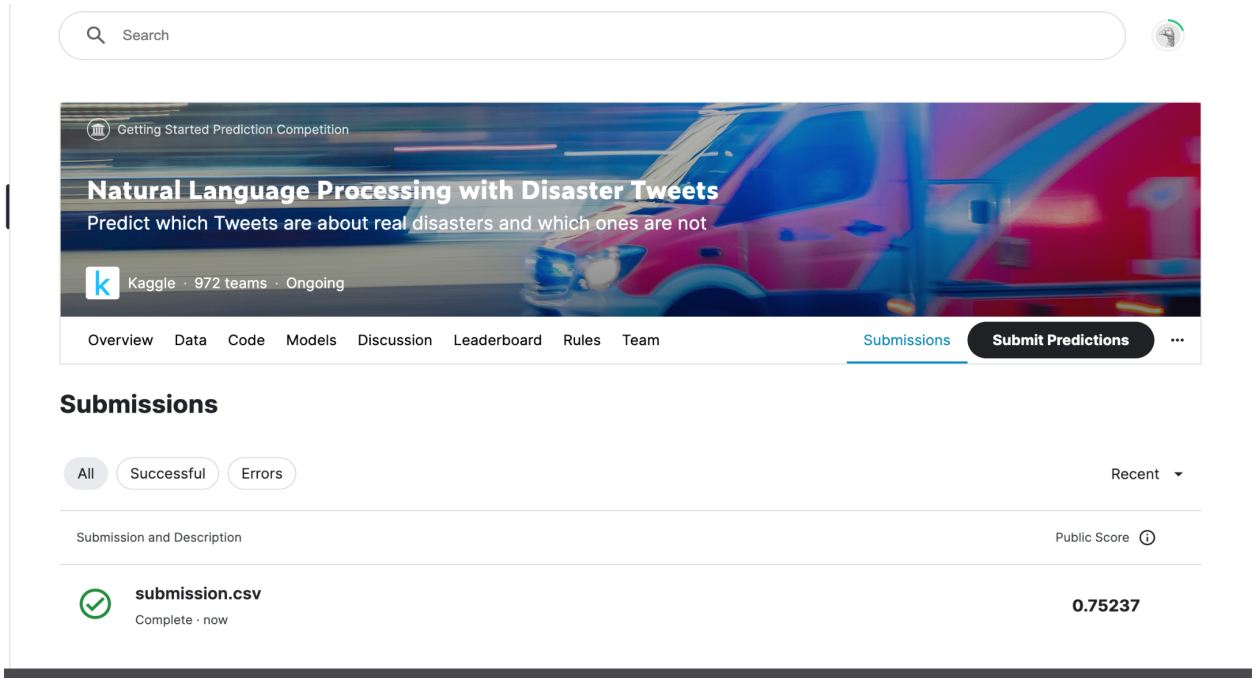
(g) Which model performed the best in predicting whether a tweet is a real disaster or not? Include your performance metric in your response. Comment on the pros and cons of using generative vs discriminative models. Think about the assumptions that Naive Bayes makes. How are the assumptions differ- ent from logistic regressions? Discuss whether it's valid and efficient to use Bernoulli Naive Bayes classifier for natural language texts.

Bernoulli Naive Bayes model (generative) we got a F-1 score of 0.7320. Logistic Regression (discriminative) gives us a F1 score of around 0.7990 with either regularization methods. Generative models aim to understand how data is distributed across the entire feature space, while discriminative models focus on defining decision boundaries within the data space. A generative model is primarily concerned with explaining the data generation process, whereas a discriminative model is primarily concerned with assigning labels or making predictions for the data. Generative models are more sensitive to outliers in the data compared to discriminative models. Finally, in terms of computational resources, discriminative models tend to be more computationally efficient than generative models. The Bernoulli Naive Bayes classifier has a drawback: when it encounters words in the test data for a specific class that were not seen during training, it can result in zero probabilities for that class. This issue is commonly addressed through a technique called smoothing, where a small smoothing factor is added to both the numerator and denominator of every probability calculation. This adjustment ensures that even for new words, probabilities are not zero. However, it's worth noting that implementing smoothing in Bernoulli Naive Bayes can introduce some inefficiency as it involves additional computational steps and demands higher computational resources.

(h) Report your results on *training* and *development* set. Do these results differ significantly from those using the bag of words model? Discuss what this implies about the task.

The F-1 score on the training data set using a logistic regression with L1 regularization is 0.6698. The F-1 score on the development data set using a logistic regression with L1 regularization is 0.6580. The scores are much lower than what we received using the bag of words model. Using two-grams, we significantly increase the dimensionality of your data. With more features, the model may require more data to generalize effectively, or it may struggle to capture relevant patterns, leading to lower performance. Some n-grams may not carry meaningful information or could introduce noise into the model.

(i) Submit your results to Kaggle, and report the resulting *F* 1-score on the test data, as reported by Kaggle. Was this lower or higher than you expected? Discuss why it might be lower or higher than your expectation.



The F-1 score reported on Kaggle is 0.752. The score is higher than I expected. It may be because we pre-processed the data quite well, stripping it off of punctuations, removing stopwords and lemmatizing the words.

WRITTEN EXERCISES DATE

① 80 Cornell students
30 of them are Master's
50 of them are Phd

Master's students (30 out of 80):
8 Master's students who bike
9 Master's students who ski
PhD students (50 out of 80)
30 PhD students who bike
12 PhD students who ski

TARGET: $y = 1$ if student is PhD
$y = 0$ if student is Master's.

Features: x_1 = a binary indicator of whether a student ~~somebody~~ bikes
x_2 = a binary indicator of whether a student skis.

~~(a)~~ $P(x_1) = \frac{8+30}{80} = \frac{38}{80}$

$P(x_2) = \frac{9+12}{80} = \frac{21}{80}$

$P(Y = y \mid X) = (x_1, x_2))$

$P(Y \mid X) = \frac{P(X \mid Y)\, P(Y)}{P(X)}$

~~$P(x_1 = 0 \mid Y = 0)$~~ $P(x_1 = 0 \mid Y = 0) = \frac{22}{30}$ $P(x_2 = 0 \mid Y = 0) = \frac{21}{30}$

~~$P(x_2$~~ $P(x_1 = 0 \mid Y = 1) = \frac{20}{50}$ $P(x_2 = 0 \mid Y = 1) = \frac{38}{50}$

$P(Y = 0) = \frac{30}{80}$ $P(Y = 1) = \frac{50}{80}$

DATE:

a) Naive Bayes classifier assumes that predictors/ features in a Naive Bayes model are conditionally independent, or unrelated to any of the other feature in a model. It also assumes that all features contribute equally to the outcome. In this context, the features of whether a student bikes (x_1) or whether a student skis (x_2) are conditionally independent i.e. observing whether a student bikes has no influence on observing whether a student skis. Both features (whether a student bikes or skis) contribute equally in determining to which class a student belongs (Master's or PhD).

b) $P(Y=0 \mid X_1=0; X_2=0) = \dfrac{P((X_1=0, X_2=0)\mid Y=0)\cdot P(Y=0)}{\text{Total probability of } P((X_1=0, X_2=0)}$

$$= \frac{P((X_1=0, X_2=0)\mid Y=0)\cdot P(Y=0)}{P((X_1=0, X_2=0)\mid Y=0)\cdot P(Y=0) + P((X_1=0, X_2=0)\mid Y=1)\cdot P(Y=1)}$$

$$= \frac{\frac{22}{30} * \frac{21}{30} * \frac{30}{80}}{\left(\frac{22}{30} * \frac{21}{30} * \frac{30}{80}\right) + \left(\frac{38}{50} * \frac{20}{50} * \frac{50}{80}\right)}$$

$$= \frac{77/400}{\frac{77}{400} + \frac{19}{100}}$$

$$= \frac{77}{153}.$$

c) Given that every PhD who skis also bikes.

$P(B \cap S)$ for PhD $= 12$.

$P(X_1=0, X_2=0 \mid Y=1) = 1 - \frac{30}{50} = \frac{20}{50}$.

Using the same formula as above (1b)

$$P(Y=0 \mid X_1=0; X_2=0) = \frac{P((X_1=0, X_2=0)\mid Y=0)\cdot P(Y=0)}{P((X_1=0, X_2=0)\mid Y=0)\cdot P(Y=0) + P((X_1=0, X_2=0)\mid Y=1)\cdot P(Y=1)}$$

$$= \frac{77/400}{\frac{77}{400} + \left(\frac{20}{50} * \frac{50}{80}\right)}$$

$$= \frac{77/400}{77/400 + \frac{100}{400}} = \frac{77}{177}.$$

$P_\theta(y=k) = \emptyset_k$.

2a) Log Likelihood estimate

$$\log P_\theta(x^i, y^i)$$

$$P_\theta(y) = \emptyset_k.$$

$$P_\theta(x|y) = \psi_{jkl}$$

$$P_\theta(x|y=k) = \prod_{j=1}^{d} P_\theta(x_j|y=k)$$

$$P(x^i, y^i) = \log P_\theta(x^i, y^i)$$

$$= \log \prod_{j=1}^{d} P(x_j, y=k)$$

Now we know,

$$P + (1-P) = 1.$$

Bernoulli's distribution states

$$P(x) = P^x(1-P)^{1-x}.$$

$$P(x_j, y=k)$$

Now, Bernoulli distribution has a single parameter P. We replace y with $P_\theta(y=k) = \emptyset_k$.

The question gives us a distribution over a vector of features x, ~~can be~~ which is.

$$P_\theta(x|y=K) = \prod_{j=1}^{d} P_\theta(x_j|y=k).$$

We factor in the probability function above with

$$\emptyset_K^{x_j}(1-\emptyset_K)^{1-x_j}.$$

Thus, we get $\prod_{j=1}^{d} \emptyset_k^{x_j}(1-\emptyset_K)^{1-x_j}$

When we add log to both sides of the equation, we get.

$$\log \prod_{j=1}^{d} \phi_k^{x_j} (1-\phi_k)^{1-x_j}$$

Due to the log operators the multiplication is correlated to summation.

$$\log \prod_{j=1}^{d} \phi_k^{x_j} (1-\phi_k)^{1-x_j} = \sum_{j=1}^{d} \log \phi_k^{x_j} + \sum_{j=1}^{d} \log (1-\phi_k)^{1-x_j}$$

~~Using rules of log power.~~

We replace j with i and d with n.

$$\log \prod_{i=1}^{n} \phi_k^{x_j} (1-\phi_k)^{1-x_j} = \sum_{i=1}^{n} \log \phi_k^{x_i} + \sum_{i=1}^{n} \log (1-\phi_k)^{1-x_i}$$

$$= \sum_{i=1}^{n} x^i \log \phi_k + \sum_{i=1}^{n} (1-x_i) \log (1-\phi_k).$$

The log elements do not depend on i/j when we bring them out of the summation.

$$\log \phi_k \sum_{i=1}^{n} x^i + \log (1-\phi_k) \sum_{i=1}^{n} (1-x_i).$$

$$= \log \phi_k \sum_{i=1}^{n} x^i + \log (1-\phi_k) \left(N - \sum_{i=1}^{n} x_i\right)$$

We take the derivative of the above at 0 to get maximum likelihood.

$$\frac{\sum_{i=1}^{n} x^i}{\phi_k^* \ln 10} - \frac{\left(N - \sum_{i=1}^{n} x_i\right)}{(1-\phi_k^*) \ln 10} = 0$$

Unifying the denominator

$$\frac{(1-\phi_k^*) \sum_{i=1}^{n} x_i - \phi_k^* \left(N - \sum_{i=1}^{n} x_i\right)}{\phi_k^* (1-\phi_k^*) \ln 10} = 0$$

DATE:

$$(1-\phi_k^*)\sum_{i=1}^{n} x^i - \phi_k^*\left(N - \sum_{i=1}^{n} x^i\right) = 0$$

$$= \sum_{i=1}^{n} x^i - \cancel{\phi_k^* \sum_{i=1}^{n} x^i} - \phi_k^* N + \cancel{\phi_k^* \sum_{i=1}^{n} x_i} = 0$$

$$\Rightarrow \phi_k^* N = \sum_{i=1}^{n} x^i$$

$$\Rightarrow \phi_k^* = \frac{\sum_{i=1}^{n} x^i}{N}$$

we know $\sum_{j=1}^{n} x^j$ is number of features for class k so we can write it as:

$$\phi_k^* = \frac{N^k}{N}.$$

b) $\ell(\Theta) = \sum_{i=y^i=k} \log P(x_j^i \mid y^i ; \psi_{jkl})$

$$= \sum_{i=y^{(i)}=k} \log \left(\frac{\psi_{jk} x_j^i}{\sum_{k=1}^{K} \psi_{jkl}} \right).$$

$$= \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \log \frac{\psi_{jk} x_j^i}{\sum_{l=1}^{L} \psi_{jkl}}$$

$$\frac{d}{d\psi_{jkl}} (\ell(\Theta)) = \frac{d}{d\psi_{jkl}} \left(\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \log \frac{\psi_{jk} x_j^i}{\sum_{l=1}^{L} \psi_{jkl}} - \right).$$

$$= \frac{d}{d\psi_{jkl}} \left(\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \log \frac{\psi_{jk} x_j^i}{\sum_{l=1}^{L} \psi_{jkl}} \right)$$

$$= \frac{d}{d\psi_{jkl}} \left(\sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{i=y^i=k} \left(\log(\psi_{jk} x_j^i) - \log\left(\sum_{l=1}^{L} \psi_{jkl} \right) \right) \right)$$

$$= \frac{n_{jkl}}{\psi_{jkl}} - n_k = 0$$

$$\Rightarrow \quad \psi_{jkl}^{*} = \frac{n_{jkl}}{n_k}$$