## assignment-7

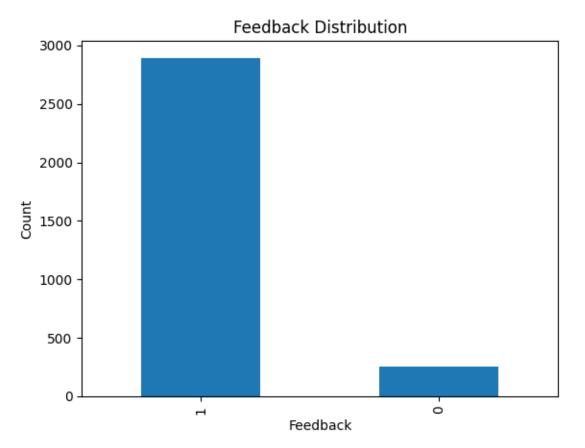
## April 17, 2024

## 0.1 Problem Statement

Consider the Amazon Alexa Reviews Dataset. This dataset consists of a nearly 3000 Amazon customer reviews (input text), star ratings, date of review, variant and feedback of various amazon Alexa products like Alexa Echo, Echo dots, Alexa Firesticks etc. Perform following operations on this dataset. \* (i) Plot a graph of Positive and Negative Feedback (1 = Positive Feedback, 0 = Negative Feedback) \* (ii) Plot the graph of Ratings distribution. \* (iii) Convert the review text into lowercase \* (iv) Remove all punctuations from review text. \* (v) Remove emoticons and emojis from the text \* (vi) Tokenize the review text into words. \* (vii) Remove the Stopwords from the tokenized text. \* (viii) Perform stemming & lemmatization on the review text. \* (ix) Perform the word vectorization on review text using Bag of Words technique. \* (x) Create representation of Review Text by calculating Term Frequency and Inverse Document Frequency (TF-IDF)

```
[2]: !pip install emoji
     Collecting emoji
       Downloading emoji-2.11.0-py2.py3-none-any.whl (433 kB)
                                 433.8/433.8
     kB 3.1 MB/s eta 0:00:00
     Installing collected packages: emoji
     Successfully installed emoji-2.11.0
 [3]: import pandas as pd
      import matplotlib.pyplot as plt
      import nltk
      from nltk.corpus import stopwords
      from nltk.tokenize import word_tokenize
      from nltk.stem import PorterStemmer, WordNetLemmatizer
      from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
      import string
      import emoji
[43]: df = pd.read csv('/content/Alexa-Dataset.csv')
[44]:
     df.isna().sum()
[44]: rating
                          0
                          0
      date
```

```
0
      variation
      verified_reviews
                          1
      feedback
                          0
      dtype: int64
[45]: df['verified_reviews'] = df['verified_reviews'].fillna('')
[46]: df.isna().sum()
[46]: rating
                          0
                          0
      date
      variation
                          0
      verified_reviews
      feedback
      dtype: int64
[47]: # (i) Plot a graph of Positive and Negative Feedback
      df['feedback'].value_counts().plot(kind='bar')
      plt.title('Feedback Distribution')
      plt.xlabel('Feedback')
      plt.ylabel('Count')
      plt.show()
```



```
[48]: # (ii) Plot the graph of Ratings distribution
df['rating'].value_counts().plot(kind='bar')
plt.title('Ratings Distribution')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.show()
```

## Ratings Distribution 2000 - 1500 - 500 - 7 Rating

```
[51]: # (iv) Remove all punctuations from review text
      df['verified reviews'] = df['verified reviews'].apply(lambda x: x.translate(str.
       →maketrans('', '', string.punctuation)))
      df['verified reviews'].head()
[51]: 0
                                                 love my echo
                                                     loved it
      1
      2
           sometimes while playing a game you can answer ...
      3
           i have had a lot of fun with this thing my 4 y...
      Name: verified reviews, dtype: object
[52]: # (v) Remove emotions and emojis from the text
      def remove_emoji(text):
          return text.encode('ascii', 'ignore').decode('ascii')
      df['verified_reviews'] = df['verified_reviews'].apply(remove_emoji)
      df['verified_reviews'].head()
[52]: 0
                                                 love my echo
                                                     loved it
           sometimes while playing a game you can answer ...
      3
           i have had a lot of fun with this thing my 4 y...
                                                        music
      Name: verified_reviews, dtype: object
[53]: import nltk
      nltk.download('punkt')
     [nltk data] Downloading package punkt to /root/nltk data...
                   Package punkt is already up-to-date!
     [nltk data]
[53]: True
[54]: # (vi) Tokenize the review text into words
      df['tokenized_text'] = df['verified_reviews'].apply(lambda x: x.split())
      df['tokenized_text']
[54]: 0
                                                [love, my, echo]
      1
                                                      [loved, it]
      2
              [sometimes, while, playing, a, game, you, can,...
              [i, have, had, a, lot, of, fun, with, this, th...
      3
                                                          [music]
              [perfect, for, kids, adults, and, everyone, in...
      3145
      3146
              [listening, to, music, searching, locations, c...
      3147
              [i, do, love, these, things, i, have, them, ru...
```

```
3148
              [only, complaint, i, have, is, that, the, soun...
      3149
                                                           [good]
      Name: tokenized_text, Length: 3150, dtype: object
[55]: import nltk
      nltk.download('stopwords')
     [nltk_data] Downloading package stopwords to /root/nltk_data...
     [nltk_data]
                   Package stopwords is already up-to-date!
[55]: True
[56]: # (vii) Remove the Stopwords from the tokenized text
      stop words = set(stopwords.words('english'))
      df['filtered_text'] = df['tokenized_text'].apply(lambda x: [word for word in x_
       →if word not in stop_words])
      df['filtered_text']
[56]: 0
                                                     [love, echo]
                                                          [loved]
      1
      2
              [sometimes, playing, game, answer, question, c...
      3
              [lot, fun, thing, 4, yr, old, learns, dinosaur...
      4
                                                          [music]
      3145
                               [perfect, kids, adults, everyone]
              [listening, music, searching, locations, check...
      3146
      3147
              [love, things, running, entire, home, tv, ligh...
      3148
              [complaint, sound, quality, isnt, great, mostl...
      3149
                                                           [good]
      Name: filtered_text, Length: 3150, dtype: object
[57]: import nltk
      nltk.download('wordnet')
     [nltk_data] Downloading package wordnet to /root/nltk_data...
     [nltk_data]
                   Package wordnet is already up-to-date!
[57]: True
[58]: # (viii) Perform stemming & lemmatization on the review text
      porter = PorterStemmer()
      wordnet_lemmatizer = WordNetLemmatizer()
      df['stemmed_text'] = df['filtered_text'].apply(lambda x: [porter.stem(word) for___
       →word in x])
      df['lemmatized_text'] = df['filtered_text'].apply(lambda x: [wordnet_lemmatizer.
       →lemmatize(word) for word in x])
```

```
[59]: df['stemmed_text']
[59]: 0
                                                      [love, echo]
                                                            [love]
      1
      2
               [sometim, play, game, answer, question, correc...
      3
               [lot, fun, thing, 4, yr, old, learn, dinosaur,...
      4
      3145
                                   [perfect, kid, adult, everyon]
      3146
               [listen, music, search, locat, check, time, lo...
               [love, thing, run, entir, home, tv, light, the...
      3147
      3148
               [complaint, sound, qualiti, isnt, great, mostl...
      3149
                                                            [good]
      Name: stemmed_text, Length: 3150, dtype: object
[60]: df['lemmatized_text']
[60]: 0
                                                      [love, echo]
      1
                                                           [loved]
      2
               [sometimes, playing, game, answer, question, c...
      3
               [lot, fun, thing, 4, yr, old, learns, dinosaur...
      4
                                                           [music]
      3145
                                  [perfect, kid, adult, everyone]
      3146
               [listening, music, searching, location, checki...
      3147
               [love, thing, running, entire, home, tv, light...
      3148
               [complaint, sound, quality, isnt, great, mostl...
      3149
                                                            [good]
      Name: lemmatized_text, Length: 3150, dtype: object
[61]: # (ix) Perform the word vectorization on review text using Bag of Words,
       \hookrightarrow technique
      vectorizer = CountVectorizer()
      X_bow = vectorizer.fit_transform(df['lemmatized_text'].apply(lambda x: ' '.
        \hookrightarrowjoin(x)))
[62]: # Convert the BoW matrix to a DataFrame
      bow_df = pd.DataFrame(X_bow.toarray(), columns=vectorizer.
       # View the BoW DataFrame
      print(bow_df)
                                                                      1220 ... \
            072318 10
                        100
                             1000
                                    100x
                                         1010
                                                1030pm
                                                             1100sf
                                                        11
                 0
                          0
                                       0
     0
                     0
                                 0
                                             0
                                                      0
                                                          0
                                                                   0
                                                                         0 ...
                 0
                     0
                          0
                                 0
                                       0
                                             0
                                                      0
                                                          0
                                                                   0
                                                                         0
     1
     2
                 0
                     0
                          0
                                 0
                                       0
                                             0
                                                      0
                                                          0
                                                                   0
                                                                         0
```

			0	0	0	0	0	0	0	0	•••	
4	0	0	0	0	0	0	0	0	0	0		
•••		•••			•••			•••				
3145	0	0	0	0	0	0	0	0	0	0		
3146	0	0	0	0	0	0	0	0	0	0	•••	
3147	0	0	0	0	0	0	0	0	0	0	•••	
3148	0	0	0	0	0	0	0	0	0	0	•••	
3149	0	0	0	0	0	0	0	0	0	0	•••	
	youtube	s yo	uve	yr yı	ıp zer	o zig	bee zo	nkedoı	ıt zwav	e zz	ZZZ	ZZZZZZ
0		0	0	0	0	0	0		0	0	0	
1		0	0	0	0	0	0		0	0	0	
2		0	0	0	0	0	0		0	0	0	
3		0	0	1	0	0	0		0	0	0	
4		0	0	0	0	0	0		0	0	0	
•••	•••			•••	•••	•••			•••			
3145		0	0	0	0	0	0		0	0	0	
3146		0	0	0	0	0	0		0	0	0	
3147		0	0	0	0	0	0		0	0	0	
3148		0	0	0	0	0	0		0	0	0	
3149		0	0	0	0	0	0		0	0	0	
$\# (x)$ $\hookrightarrow In$	rows x  Create verse Do	repr	resent nt Fre	ation equenc	y (TF-1	IDF)	xt by c	alcula	ting Ter	rm Fr	reque	ncy an
# (x)  → In  tfidf  X_tfi  →'	Create verse Do Lote '.join(x	reprocumentizer idf_v	resent nt Fre = Tfi ector	ation equenc dfVect izer.f	y (TF-1 corizer it_tra	() nsform	(df['le		-			
# (x)  In  tfidf  X_tfi  Con  tfidf  # Con  tfidf  Yei	Create verse Do Lyector ddf = tf	reprocument izer idf_vel()))  e TF-d.Datre_nar	resent nt Fre = Tfi ector IDF maFram	ation equencedfVect izer.f eatrix e(X_tf	y (TF-1 corizer fit_tra	() nsform $ataFra$	(df['le	mmatiz	ed_text	'].ap	ply(	
# (x)  In  tfidf  X_tfi  Con  tfidf  # Con  tfidf  Yei	Create verse Do vector df = tf '.join(x  vert th df = p t_featur  wert th	reprocument izer idf_vel()))  e TF-d.Datre_nar	resent nt Fre = Tfi ector IDF maFram	ation equencedfVect izer.f eatrix e(X_tf	y (TF-1 corizer fit_tra	() nsform $ataFra$	(df['le	mmatiz mns=tf	ed_text	'].ap	ply(	lambda
# (x)  In  tfidf  X_tfi  Con  tfidf  # Con  tfidf  Yei	Create verse Do	reprocument izer idf_velt))  e TF-d.Date_nar  F-IDF	resent nt Fre = Tfi ector  TIDF m aFram nes_ou	ation equence dfVect izer.f eatrix e(X_tf it()) Frame	y (TF-) corizer cit_tra to a D cidf.to	TDF) () nsform ataFra array(	(df['le	mmatiz mns=tf	ed_text	].ap	ply(	lambda
# (x)  In thick this is the second thick	Create verse Do Evector df = tf '.join(x  vert th Edf = pe t_featur ew the Te c(tfidf_c	reprocument izer idf_vel()))  e TF-d.Date_nar  F-IDFdf)	resent nt Fre = Tfi ector  IDF m aFram nes_ou Data	ation equence dfVect izer.f  atrix e(X_tf it())  Frame	y (TF-) corizer fit_tra  to a D fidf.to	TDF) () msform  ataFra array(	(df['le: me ), colu	mmatiz	ed_text idf_vect	122 0.	ply( er.	lambda
# (x)  In thick  thick  # Continue  thick  # Vie  print  0 1	Create verse Do Evector idf = tf '.join(x  vert th Edf = pe t_featur  c(tfidf_c  072318 0.0	reprocument izer idf_veltilities idf_veltiliti	resent nt Fre = Tfi ector  IDF maFram nes_ov  Data  100 0.0	ation equence dfVect izer.f  atrix e(X_tf it())  Frame  1000 0.0	y (TF-) corizer fit_tra  to a D fidf.to  100x 0.0	TDF) () msform  ataFra array(  1010 0.0	(df['le: me ), colu	mmatiz	ed_text idf_vect 1100sf 0.0	122 0. 0.	er.	lambda
# (x)  tfidf  X_tfi  # Con  tfidf  sget  # Vie  print  0 1 2	Create verse Do E_vector idf = tf '.join(x  evert th E_df = pe t_featur  c(tfidf_e)  072318  0.0  0.0	reprocument izer idf_velt))  e TF-d.Date_nar  F-IDF df)  10 0.0 0.0	resent nt Fre = Tfi ector  TDF m aFram nes_ou  100 0.0 0.0	ation equence dfVect izer.f eatrix e(X_tf it()) Frame  1000 0.0 0.0	y (TF-) corizer fit_tra  to a D fidf.to  100x 0.0 0.0	() msform  ataFra array(  1010 0.0 0.0	(df['le: me ), colu	mmatiz mns=tf	ed_text idf_vect 1100sf 0.0 0.0	122 0. 0.	er.	lambda
# (x)  tfidf  X_tfi  Y Con  tfidf  tfidf  perint  0 1 2 3	Create verse Do Evector df = tf '.join(x  vert th Edf = p t_featur  c(tfidf_c  072318 0.0 0.0 0.0	reprocument izer idf_velt))  e TF-d.Date_nar  F-IDFdf)  10 0.0 0.0 0.0	resent nt Fre = Tfi ector  IDF m aFram nes_ou  100 0.0 0.0 0.0	ation equence dfVect izer.f  atrix e(X_tf it())  Frame  1000 0.0 0.0 0.0	y (TF-) corizer fit_tra  to a D fidf.to  100x 0.0 0.0 0.0	() nsform  ataFra array(  1010 0.0 0.0 0.0	(df['le: me ), colum 0.0 0.0 0.0	mmatiz mns=tf	ed_text  idf_vect  1100sf 0.0 0.0 0.0	122 0. 0. 0.	er.  20 0 0 0	lambda . \
# (x)  tfidf  X_tfi  # Con  tfidf  get  # Vie  print  0 1 2 3 4	Create verse Do Evector df = tf '.join(x  vert th Edf = pe t_featur  (tfidf_c)  072318  0.0  0.0  0.0  0.0  0.0	reprocument izer idf_velo))  e TF-d.Datre_nar  F-IDF df)  10 0.0 0.0 0.0 0.0	resent nt Free = Tfi ector  IDF m aFram nes_ou  100 0.0 0.0 0.0 0.0 0.0	ation equence dfVect izer.f  atrix e(X_tf it())  Frame  1000 0.0 0.0 0.0 0.0	y (TF-) corizer fit_tra  to a D fidf.to  100x 0.0 0.0 0.0 0.0	1010 0.0 0.0 0.0 0.0	(df['le: me ), colum 0.0 0.0 0.0 0.0	mmatiz mns=tf	idf_vect	122 0. 0. 0.	er.  20 0 0 0	lambda . \
# (x)  tfidf  X_tfi  Y Con  tfidf  sget  # Vie  print	Create verse Do Evector df = tf '.join(x  vert th Edf = pe t_featur  (tfidf_c)  072318  0.0  0.0  0.0  0.0  0.0	reprocument izer idf_vel))  e TF-d.Date_nar  F-IDF df)  10 0.0 0.0 0.0 0.0 0.0	resent nt Fre = Tfi ector  IDF m aFram nes_ou  O.0 O.0 O.0 O.0 O.0 O.0 O.0 O.0 O.0 O.	ation equence dfVect izer.f  atrix e(X_tf it())  Frame  1000 0.0 0.0 0.0 0.0 0.0	100x 0.0 0.0 0.0 0.0	1010 0.0 0.0 0.0 0.0	(df['le: me ), colum 0.0 0.0 0.0 0.0	mmatiz mns=tf	idf_vect	122 0. 0. 0.	er.	lambda . \
# (x) tfidf X_tfidf X_tfidf  # Continue tfidf  # Vie print  0 1 2 3 4 3145	Create verse Do Evector df = tf '.join(x  vert th Edf = p t_featur  c(tfidf_ 072318 0.0 0.0 0.0 0.0 0.0	reprocument izer idf_velt))  e TF-d.Date_nar  F-IDFdf)  10 0.0 0.0 0.0 0.0 0.0	resent nt Fre = Tfi ector  IDF m aFram nes_ou  O.0 O.0 O.0 O.0 O.0 O.0 O.0 O.0 O.0 O.	ation equence dfVect izer.f  atrix e(X_tf it())  Frame  1000 0.0 0.0 0.0 0.0	100x 0.0 0.0 0.0 0.0	1010 0.0 0.0 0.0	(df['le: me ), colum 0.0 0.0 0.0 0.0	mmatiz mns=tf	1100sf 0.0 0.0 0.0	122 0. 0. 0. 0.	er.	lambda . \
# (x)  tfidf  X_tfi  # Con  tfidf  get  # Vie  print  0 1 2 3 4	Create verse Do E_vector df = tf '.join(x  vert th E_df = po t_featur  c(tfidf_c  072318 0.0 0.0 0.0 0.0 0.0 0.0 0.0	reprocument izer idf_vel()))  e TF-d.Date_name F-IDF df)  10 0.0 0.0 0.0 0.0 0.0 0.0	resent nt Fre = Tfi ector  IDF m aFram nes_ou  O.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0	ation equence dfVect izer.f  atrix e(X_tf it())  Frame  1000 0.0 0.0 0.0 0.0 0.0 0.0	100x 0.0 0.0 0.0 0.0 0.0	1010 0.0 0.0 0.0 0.0	1030pm 0.0 0.0 0.0 0.0 0.0 0.0	mns=tf	1100sf 0.0 0.0 0.0 0.0	122 0. 0. 0. 0.	er.  20 0 0 0 0	lambda . \

```
youtubes youve yr yup zero zigbee zonkedout zwave zzzz \
0
         0.0
              0.0 0.000000 0.0
                                0.0
                                      0.0
                                               0.0
                                                     0.0
                                                          0.0
        0.0
              0.0 0.000000 0.0
                                0.0
                                      0.0
                                               0.0
                                                     0.0
                                                          0.0
1
2
        0.0
                                               0.0
                                                     0.0
                                                          0.0
              0.0 0.000000
                           0.0
                                0.0
                                      0.0
        0.0
              0.0 0.327532
                                               0.0
                                                     0.0
3
                           0.0
                                0.0
                                      0.0
                                                          0.0
         0.0
              0.0 0.000000
4
                           0.0
                                0.0
                                      0.0
                                               0.0
                                                     0.0
                                                          0.0
                                      •••
                 ... ... ...
                           0.0
                                0.0
                                               0.0
                                                     0.0
                                                          0.0
3145
       0.0
              0.0 0.000000
                                      0.0
       0.0
              0.0 0.000000
                                0.0
                                               0.0
                                                     0.0
                                                          0.0
3146
                           0.0
                                      0.0
                                0.0
3147
       0.0
              0.0 0.000000
                           0.0
                                      0.0
                                               0.0
                                                     0.0
                                                          0.0
                                0.0
                                      0.0
        0.0
              0.0 0.000000
                                               0.0
                                                     0.0
                                                          0.0
3148
                           0.0
3149
        0.0
              0.0 0.000000 0.0
                                               0.0
                                                     0.0
                                                          0.0
    ZZZZZZZ
0
        0.0
        0.0
1
2
        0.0
3
        0.0
4
        0.0
•••
        0.0
3145
3146
        0.0
3147
        0.0
3148
        0.0
3149
        0.0
```

[3150 rows x 3999 columns]

[]: