Data Driven - Assignment 2

Azamat Shora

September 21, 2023

1 Describe Yelp data.

First of all, we uploaded our data and converted it into the data frame. With a shape of (1000, 2). Which means we have 2 attributes and 1000 rows.

The first column is responsible for the text, its length in this dataset can jump between 11 characters up to 149 characters.

The second column contains only numbers 0 or 1, and it is responsible for the class, which is determined by the text from the previous column.

Link to the code : **link**

	class
count	1000.00000
mean	0.50000
std	0.50025
min	0.00000
25%	0.00000
50%	0.50000
75%	1.00000
max	1.00000

Figure 1: Describe data of Yelp file.

The reference of this file is next : "Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." Advances in neural information processing systems 28 (2015)."

2 Process SentiWord data, describe the result.

The references: "Gatti, L., Guerini, M., Turchi, M. (2015). SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. IEEE Transactions on Affective Computing, 7(4), 409- 421."

In this document, we used this file with the following **link**

The first column is responsible for the text, the text itself consists of a set of characters or letters, followed by a '' sign and one of these letters ('n', 'a', 'r', 'v'). In this case, n is a noun, a is an adjective, r is an adverb, and v is a verb.

The second column contains only numbers from -1 to 1, and it is responsible for 'polarity' which is determined by the text from the previous column.

For convenience in further work with this dataframe, the first column was divided by the delimiter ''. The new first column contained everything before '' and is called 'lemma'. And the second character that was right after '' is one of these ('n', 'a', 'r', 'v').

The smallest value of "polarity" in this table was -0.93489, and the largest was 0.89489. The table dimensions themselves became (155287, 3).

In order to better understand which words or expressions appeared more in a given dataset by group of ('n', 'a', 'r', 'v') I built a table where I showed the average polarity value.

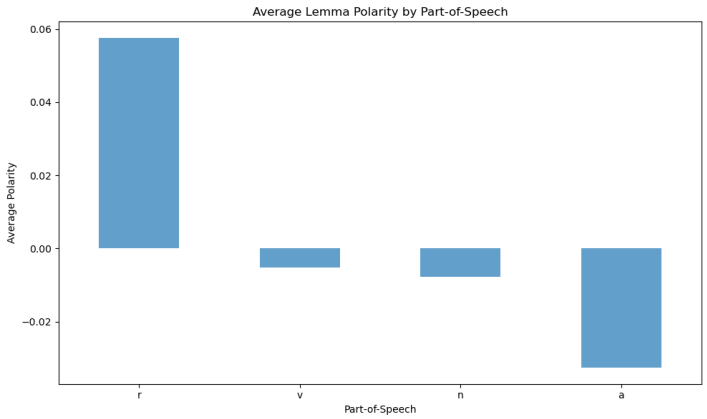

Figure 2: Average Lemma Polarity by Part-of-Speech.

Table 1: Data Summary

2*Properties	Values				
	N	A	V	R	Total
Number of Positive	9271	5322	2299	2201	117798
Number of Negative	12393	7325	2278	513	21479
Number of Neutral	96134	8832	6952	1767	11529
Number of Average	-0.0077	-0.0325	-0.0052	0.0575	
Number of Variance	0.0102	0.0437	0.0266	0.0165	
Number of Std. Deviation	0.1011	0.2092	0.1632	0.1284	

3 Develop a lexicon-based sentiment classifier using Stanza for lemmatization and POS-tagging. (Mind difference in labeling: sentences: 0 – negative, 1 – positive; words: continuous scores from the range [-1, 1]. Note that SentiWords and Stanza use different POS tag sets.)

I then for each of the 4 choices ('Adjectives', 'Nouns', 'Verbs', 'Adverbs') Show how many 'Positive', 'Negative' and 'Neutral' amounts of 'polarity' appear in percentages and quantities.

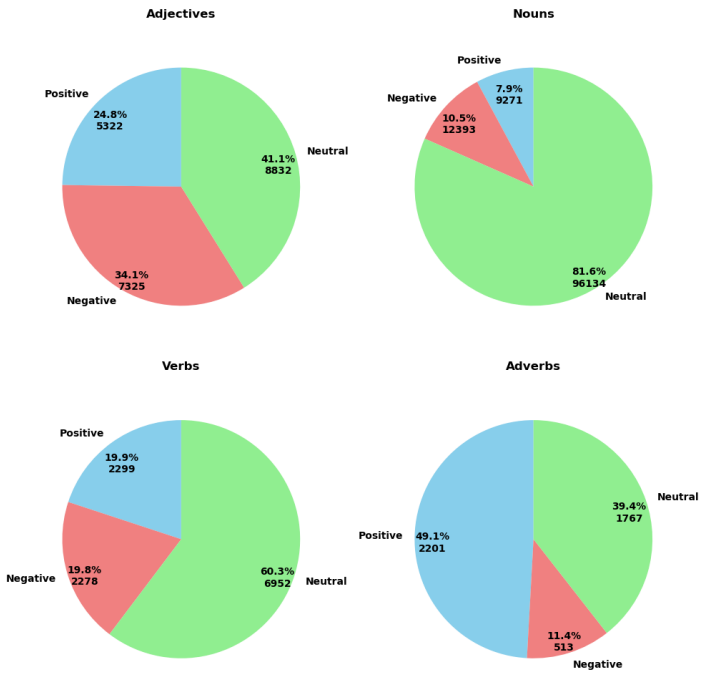

Figure 3: Percentages of positive/Neutral/Negative for each field in ('Adjectives', 'Nouns', 'Verbs', 'Adverbs').

4 Evaluate the classifier on the Yelp data, report confusion matrix and F1 scores for each class (negative and positive).

I then created a new column in the yelp dataframe called 'polarity'. And after that, I built a ConfusionMatrix for yelp['class'] and yelp['polarity']. The results are as follows:

Table 2: F1-Scores for Yelp

F1-Score (Negative)	0.3262
F1-Score (Positive)	0.7064

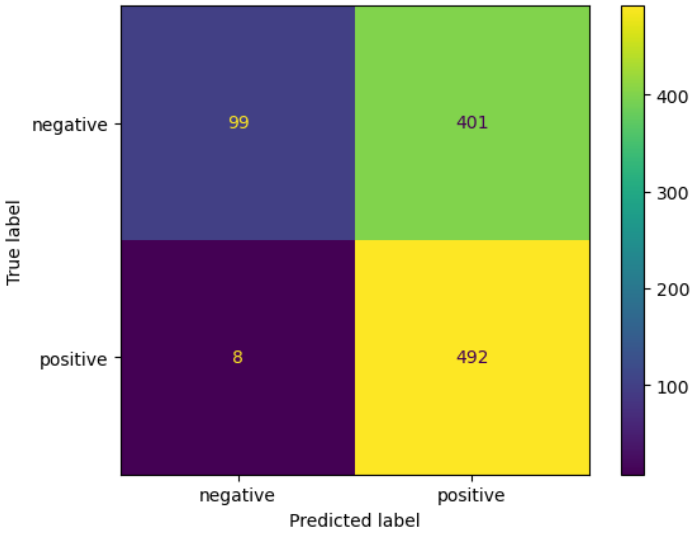

Figure 4: Confusion matrix for Yelp data

5 Apply Stanza's sentiment analysis model to the Yelp data, report confusion matrix and F1 scores for each class (negative and positive).

I then created a new column in the yelp dataframe called 'stanza$_p olarity'. And after that I built a Confusion Matrix for yelp['clas

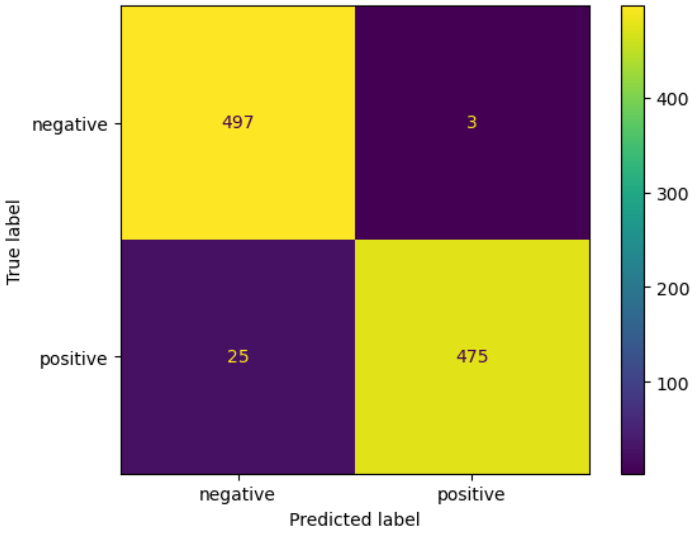

Figure 5: Confusion matrix for Yelp data

Table 3: F1-Scores for Yelp

F1-Score (Negative)	0.9726
F1-Score (Positive)	0.9714