Sentiment analysis

Yelp has 1000 sentences and their corresponding binary sentiment scores (0/1). The dataset was split into train, validation, and test sets with 70/10/20 ratio. I vectorized yelp text and trained MultinomialNB and SVC, picked the best configurations on validation set, and tested them. Then I created my own classifier and picked the best threshold on validation set, then tested. Trying out combinations of parameters yielded the best configurations for:
Vectorizer - minimum difference of 2, stopwords: None, binary: False, lowercase: True.
MultinomialNB - alpha: 1.0.
SVC - kernel: sigmoid.
My classifier - threshold: 0.1.
I reused the same vectorizer for both SVC and MultinomialNB.
Using MultinomialNB, the highest accuracy on validation set was 0.85 (random state = 0), and on test set it was 0.82.
SVC's best accuracy was 0.84 on validation set, and 0.795 on test set.
In my classifier, the best accuracy was 0.77 on validation set, and then further 0.73 on test set.

- Following words influence other words, mostly ones that follow, so I took them into account and used as a simple multiplier, for example negation being -1.
  negations = "not", "never", "no" intensifiers = "very", "extremely", "highly" diminishers = "somewhat", "slightly", "barely" For instance, "not" as a separate word appeared 116 times, "barely" 5 times, and "very" 103 times. Accuracy before the change was 0.6.

- Other influencing words could be taken into account, but I decided not to peek into the data itself, not until I have implemented what made sense in theory. So I also chose words mentioned above by thinking what would be sensible, not what is in the dataset.

- 28 lemmas were not present in sentiWords, and 686 lemmas could not be identified by Stanza, which is a loss of information from the dataset.

- SVC and MultinomialNB have very close precisions between the classes of 0 and 1, while mine is relatively worse on identifying positive sentiment (0.68) than negative (0.83). In practice, they had more data as vectorization works with any lemma. Overall, the classifier I have created is very primitive, yet close to SCV and MultinomialNB by accuracy.

- My classifier, unlike others, scored "Their menu is diverse, and reasonably priced" negative, worth noting that if no word is found in SentiWord, it gives a score of negative.

- "This is was due to the fact that it took 20 minutes to be acknowledged, then another 35 minutes to get our food...and they kept forgetting things" was misclassified only by my classifier as positive.

- However, only my classifier scored "That said, our mouths and bellies were still quite pleased" as positive, which is what ground truth is. Same was true occasionally with other sentences.

Class	Precision	Recall	F1-Score	Support
0	0.81	0.84	0.82	100
1	0.83	0.80	0.82	100
Accuracy:			0.82	200
Macro avg:	0.82	0.82	0.82	200
Weighted avg:	0.82	0.82	0.82	200

Table 1: MultinomialNB's Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.80	0.79	0.79	100
1	0.79	0.80	0.80	100
Accuracy:			0.80	200
Macro avg:	0.80	0.80	0.79	200
Weighted avg:	0.80	0.80	0.79	200

Table 2: SVC's Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.83	0.58	0.68	100
1	0.68	0.88	0.77	100
Accuracy:			0.73	200
Macro avg:	0.75	0.73	0.72	200
Weighted avg:	0.75	0.73	0.72	200

Table 3: My classifier's Classification Report

Sentence	GT	MNB	SVC	MC
The Heart Attack Grill in downtown Vegas is an absolutely flat-lined excuse for a restaurant.	0	1	1	0
As always the evening was wonderful and the food delicious!	1	1	1	1
Wow very spicy but delicious.	1	1	1	1
Definitely worth venturing off the strip for the pork belly, will return next time I'm in Vegas.	1	0	1	0
Their menu is diverse, and reasonably priced.	1	1	1	0
All in all, I can assure you I'll be back.	1	0	1	1
The goat taco didn't skimp on the meat and wow what FLAVOR!	1	0	1	1
That said, our mouths and bellies were still quite pleased.	1	0	0	1
It lacked flavor, seemed undercooked, and dry.	0	0	0	0
If you want to wait for mediocre food and downright terrible service, then this is the place for you.	0	0	0	0
The ambience is wonderful and there is music playing.	1	1	1	1
Will not be back.	0	0	0	0
I didn't know pulled pork could be soooo delicious.	1	0	0	1
I promise they won't disappoint.	1	0	0	0
I've had better, not only from dedicated boba tea spots, but even from Jenni Pho.	0	0	0	1
This is was due to the fact that it took 20 minutes to be acknowledged, then another 35 minutes to get our food...and they kept forgetting things.	0	0	0	1
They have a really nice atmosphere.	1	1	1	1
Crostini that came with the salad was stale.	0	1	0	0
This place is way too overpriced for mediocre food.	0	0	0	0
Bacon is hella salty.	1	1	1	0

Table 4: Classifiers' comparison (GT - Ground Truth, MNB - MultinomialNB, MC - My Classifier)