CIS 668 Assignment 2 Report Lichen Liang

Overview of Process

- 1. Extract only the reviews from baby.txt
- 2. Tokenize into sentences and change to lower case
- 3. Get the sentence polarity from library and create the word list
- 4. Get the stop-words and negation words
- 5. For subjectivity lexicon features, remove the stop-words and Naïve Bayes Classifier to train and find the accuracy
- 6. For negation word features, remove the stop-words that are not in the negation words and Naïve Bayes Classifier to train and find the accuracy
- 7. For each feature used, classify the tokenized reviews into positive and negative and save to the file.

Process in Details

1. In this part, I split the lines in the baby.txt and for each line starting with 'reviewText', I extracted the review into a new file 'new.txt'. There are ~160k reviews extracted.

Out[3]: "Perfect for new parents. We were able to keep track of baby's feeding, sleep and diaper change schedule for the first two and a half months of her life. Made life easier when the doctor would ask questions about habits because we had it all right ther e!\nThis book is such a life saver. It has been so helpful to be able to go back to track trends, answer pediatrician questions, or communicate with each other when you are up at different times of the night with a newborn. I think it is one of those things that everyone should be required to have before they leave the hospital. We went through all the pages of the newborn version, then moved to the infant version, and will finish up the second infant book (third total) right as our baby turns 1. See other things that are must haves for baby at [...]\nHelps me know exactly how my babies day has gone with my mother in law watching him while I go to work. It also has a section for her to write notes and let me know anything she may need. "

2. For each review in the new.txt, I tokenized them into individual sentences. There are around 900k sentences.

```
tokenize_sentence = tokenize.sent_tokenize(text2)
tokenize_sentence = [w.lower() for w in tokenize_sentence]
print(len(tokenize_sentence))
print(ten(tokenize_sentence))
print(tokenize_sentence[:20])

893984
['perfect for new parents.', "we were able to keep track of baby's feeding, sleep and diaper change schedule for the first two and a half months of her life.", 'made life easier when the doctor would ask questions about habits because we had it all right there!', 'this book is such a life saver.', 'it has been so helpful to be able to go back to track trends, answer pediatrici an questions, or communicate with each other when you are up at different times of the night with a newborn.', 'i think it is one of those things that everyone should be required to have before they leave the hospital.', 'we went through all the pages of the newborn version, then moved to the infant version, and will finish up the second infant book (third total) right as our baby turns l.', 'see other things that are must haves for baby at [...]\whelps me know exactly how my babies day has gone with my mother in law watching him while i go to work.', 'it also has a section for her to write notes and let me know anything she may need.', "i couldn't be happier with this book.", 'i bought this a few times for my older son and have bought it again for my newborn.', 'this is super easy to use and helps me keep track of his daily routine.', 'when he started going to the sitter when i went back to work, it helped me know how his day went to better prepare me for how the evening would most likely go.', 'when he was sick, it help me keep track of how many diapers a day he was producing to make sure he was getting dehydrated.', 'the note sections to the side and bottom are useful too because his sitter writes in small notes about whether or not he like d his lunch or if the playtime included going for a walk, etc. excellent for moms who are wanting to keep track of their kids d aily routine even though they are at work.', 'excellent for dads to
```

3. Like the lab, get the polarity sentences, randomize them and create the words list. The list contains around 190k words.

```
sentences = sentence_polarity.sents()
                                                         [(sent, cat) for cat in sentence_polarity.categories() for sent in sentence polarity.sents(categories=cat)]
                         documents :
                         random. shuffle (documents)
                         documents[:20]
   Out[5]: [(['makes',
                                'even',
'the',
                                    claustrophobic'.
                                    on-board',
                                    quarters',
                                    seem',
                                   'fun'.
                                  pos').
                             (['feels', 'strangely', 'hollow', 'at', 'its', 'emotional', 'core', '.'],
                                 neg'),
                             (['another',
                                    of'.
                                    those',
                                    estrogen'
                                    overdose'
                                    movies',
In [6]: #set all the words, make sure they are words and change all to lower case
                         all_words_list = [word for (sent, cat) in documents for word in sent]
                         all_words = nltk.FreqDist(all_words_list)
In [7]: all_words_list = [w.lower() for w in all_words_list]
all_words_list = [w for w in all_words_list if w.isalpha()]
                         print(len(all_words list))
                        print(all_words_list[:200])
                       187486 ['makes', 'even', 'the', 'claustrophobic', 'quarters', 'seem', 'fun', 'feels', 'strangely', 'hollow', 'at', 'its', 'emotiona l', 'core', 'another', 'one', 'of', 'those', 'estrogen', 'overdose', 'movies', 'like', 'divine', 'secrets', 'of', 'the', 'ya', 'ya', 'sisterhood', 'except', 'that', 'the', 'writing', 'acting', 'and', 'character', 'development', 'are', 'a', 'lot', 'better', 'by', 'the', 'time', 'it', 'ends', 'in', 'a', 'rush', 'of', 'sequins', 'flashbulbs', 'blaring', 'brass', 'and', 'babes', 'it', 'has', 'said', 'plenty', 'about', 'how', 'show', 'business', 'has', 'infiltrated', 'every', 'corner', 'of', 'society', 'and', 'not', 'always', 'for', 'the', 'better', 'strident', 'and', 'inelegant', 'in', 'its', 'posturing', 'by', 'no', 'means', 'a', 'and', 'sure', 'to', 'ultimately', 'disappoint', 'the', 'action', 'fans', 'who', 'will', 'be', 'moved', 'to', 'the', 'edg e', 'of', 'their', 'seats', 'by', 'the', 'dynamic', 'first', 'act', 'it', 'still', 'comes', 'off', 'as', 'a', 'touching', 'tra nscendent', 'love', 'story', 'somewhat', 'clumsy', 'and', 'too', 'lethargically', 'paced', 'but', 'its', 'story', 'about', 'a', 'mysterious', 'creature', with', 'psychic', 'abilities', 'offers', 'a', 'solid', 'a', 'terrific', 'climax', 'and', 'son e', 'nice', 'chills', 'alnost', 'entirely', 'from', 'point', 'of', 'view', 'an', 'entertaining', 'if', 'somewhat', 'standardized', 'action', 'movie', 'one', 'long', 'bore', 'the', 'result', 'is', 'scary', 'parts', 'in', 'occur', 'while', 'waiting', 'for', 'things']
                         187486
```

4. Get the stop-words and negation words. We want to make sure the stop-words are not included in further processing.

```
In [8]: #get stopwords and negation words

stopwords = nltk.corpus.stopwords.words('english')
morestopwords = ['could','would','might','must','need','sha','wo','y',"'s","'d","'ll","'t","'m","'re","'ve"]
negationwords = ['no', 'not', 'never', 'none', 'nowhere', 'nothing', 'noone', 'rather', 'hardly', 'scarcely', 'rarely', 'seldom
negationwords.extend(['ain', 'aren', 'couldn', 'didn', 'doesn', 'hadn', 'hasn', 'haven', 'isn', 'ma', 'mightn', 'mustn', 'needn'
stopwords = stopwords + morestopwords

#stopwords for NOT feature should not contain negation words
stopwords2 = [w for w in stopwords if w not in negationwords]
```

5. I used subjectivity lexicon for the first feature set. First, remove the stop words, sort by frequency, and use the most common 2000 words. Then use the Naïve Bayes Classifier to train and find the accuracy.

```
In [9]: #for SL features, remove all the stopwords first
                  #calculate the classifier accuracy using Naive Bayes Classifier
                  all words = [w for w in all words list if not w in stopwords]
                 all_words = nltk.FreqDist(all_words_list)
word_item = all_words.most_common(2000)
                  word_features = [word for (word, count) in word_item]
                 print (word_features[:50])
                  ['the', 'a', 'and', 'of', 'to', 'is', 'in', 'that', 'it', 'as', 'but', 'with', 'film', 'this', 'for', 'its', 'an', 'movie', 'b e', 'on', 'you', 'not', 'by', 'about', 'one', 'more', 'like', 'has', 'are', 'at', 'from', 'than', 'all', 'his', 'have', 'so', 'if', 'or', 'story', 'i', 'too', 'just', 'who', 'into', 'what', 'most', 'out', 'no', 'much', 'even']
isStemmed = False
                              sldict[word] = [strength, posTag, isStemmed, polarity]
                        return sldict
                 SL = readSubjectivity(SLpath)
 In [11]: def SL_features(document, word_features, SL):
                       document_words = set(document)
features = {}
                       readures - II for word in word_features:
    features['V_i]'.format(word)] = (word in document_words)
# count variables for the 4 classes of subjectivity
weakPos = 0
                       strongPos = 0
weakNeg = 0
                       strongNeg = 0

for word in document_words:
                              if word in SL:
                                    strength, posTag, isStemmed, polarity = SL[word]
                                    if strength == 'weaksubj' and polarity == 'positive':
    weakPos += 1
if strength == 'strongsubj' and polarity == 'positive':
                                    if strength == 'strongsubj' and polarity == 'positive':
    strongPos += 1
if strength == 'weaksubj' and polarity == 'negative':
    weakNeg += 1
if strength == 'strongsubj' and polarity == 'negative':
    strongNeg += 1
                                    features['positivecount'] = weakPos + (2 * strongPos)
features['negativecount'] = weakNeg + (2 * strongNeg)
                SL_featuresets = [(SL_features(d, word_features, SL), c) for (d, c) in documents] train_set, test_set = SL_featuresets[1000:], SL_featuresets[:1000] classifier = nltk.NaiveBayesClassifier.train(train_set)
                 nltk.classify.accuracy(classifier, test set)
  Out[11]: 0.764
```

6. I used negation word feature for the second feature. Different from SL feature, the stop-words removed should not contain negation words as they are needed for the process. Then use the Naïve Bayes Classifier to train and find the accuracy.

Out[12]: 0.78

7. For each feature used and for each review, classify them as positive or negative. I created four text files to save the results. Since there are too many sentences and words to be processed and it takes hours to complete. I force stopped the program and will only analyse the data in the output file so far.

```
In [13]: #loop through each sentence, decide whether it's positive or negative,
            #and save this sentence to the list and its corresponding file.
#Since it takes too long due to large amount of data, I have manually stopped it and
            #only analyse the sentences categorized in the text files
            SLpositive = []
SLnegative = []
SLp = open('SLpositive.txt', 'w')
SLn = open('SLnegative.txt', 'w')
            for sentence in tokenize sentence
                  word = nltk.word_tokenize(sentence)
                 feature = SL features (word, word features, SL)
                 if classifier.classify(feature)
                     SLpositive.append(sentence)
                 SLp.write(sentence + '\n')
elif classifier.classify(feature) == 'neg':
                     SLnegative.append(sentence)
SLn.write(sentence + '\n')
            SLp.close()
            SLn. close()
            KeyboardInterrupt
                                                                 Traceback (most recent call last)
            <ipython-input-13-f693da167ff2> in <moorphise</pre>
                               SLpositive. append (sentence)
                  15
                            SLp. write(sentence + '\n')
elif classifier. classify(feature) == 'neg':
               -> 17
                              SLnegative.append(sentence)
            d:\users\owner\appdata\local\programs\python\9y\lib\site-packages\nltk\classify\naivebayes.py in classify
            (self, featureset)
                          def classify(self, featureset)
            ---> 89
                                  return self.prob_classify(featureset).max()
                          def prob classify(self, featureset):
                  91
            d:\users\owner\appdata\local\programs\python\python39\lib\site-packages\nltk\classify\naivebayes.py in prob_clas
             sify(self, featureset)
                                   for (fname, fval) in featureset.items()
                                        if (label, fname) in self._feature_probdist:
    feature_probs = self._feature_probdist[label, fname]
    logprob[label] += feature_probs.logprob(fval)
                 115
            KeyboardInterrupt:
In [14]: print(len(SLpositive))
            print (SLpositive [: 10])
            print(len(SLnegative))
            print(SLnegative[:10])
            27240
```

27240
['perfect for new parents.', "we were able to keep track of baby's feeding, sleep and diaper change schedule for the first two and a half months of her life.", 'this book is such a life saver.', 'it has been so helpful to be able to go back to track tre nds, answer pediatrician questions, or communicate with each other when you are up at different times of the night with a newb orn.', 'this is super easy to use and helps me keep track of his daily routine.', 'excellent for dads to keep track as my husb and can quickly forget what time he fed our son.', "it's also nice to be able to look back on previous days and weeks for eating and sleeping patterns.", 'my 3 month old son spend half of his days with my mother and half with a neighbor while i worke d.', 'it was the best way to have some cohesion in his life while i was at work.', 'this book is perfect!'] 56058

obuses ['made life easier when the doctor would ask questions about habits because we had it all right there!', 'i think it is one of those things that everyone should be required to have before they leave the hospital.', 'we went through all the pages of the newborn version, then moved to the infant version, and will finish up the second infant book (third total) right as our baby turns 1.', 'see other things that are must haves for baby at [...]\nhelps me know exactly how my babies day has gone with my mo there in law watching him while i go to work.', 'it also has a section for her to write notes and let me know anything she may need.', "i couldn't be happier with this book.", 'i bought this a few times for my older son and have bought it again for my newborn.', 'when he started going to the sitter when i went back to work it beload as bear to be. whorn, , when he started going to the sitter when i went back to work, it helped me know how his day went to better prepare me for how the evening would most likely go.', 'when he was sick, it help me keep track of how many diapers a day he was producing to make sure he was getting dehydrated.', 'the note sections to the side and bottom are useful too because his sitter wri in small notes about whether or not he liked his lunch or if the playtime included going for a walk, etc.excellent for mom s who are wanting to keep track of their kids daily routine even though they are at work.']

```
In [15]: #loop through each sentence decide whether it's positive or negative,
    #and save this sentence to the list and its corresponding file.
    #Since it takes too long due to large amount of data. I have manually stopped it and
    #only analyse the sentences categorized in the text files

NOIpositive = []
NOIpe = []
NOIP = open('NOIpositive.txt', 'w')
NOIn = open('NOInegative.txt', 'w')

for sentence in tokenize_sentence:
    word = nltk.word.tokenize(sentence)
    feature = NOI_features(word, word.features2, negationwords)
    if classifier2.classify(feature) == 'pos':
        NOIpositive.append(sentence)
        NOIP.write(sentence + '\n')
    elif classifier2.classify(feature) == 'neg':
        NOInegative.append(sentence)
        NOIn.write(sentence + '\n')

NOID.close()
NOID.close()
```

```
KeyboardInterrupt
                                             Traceback (most recent call last)
<ipython-input-15-9b5afb424ba7> in <module>
12 word = nltk.word_tokenize(sentence)
            feature = NOT_features(word, word_features2, negationwords)
  -> 14
              if classifier2. classify(feature) ==
                                                    'pos'
                NOTpositive.append(sentence)
     16
                NOTp. write (sentence +
d:\users\owner\appdata\local\programs\python\python39\lib\site-packages\nltk\classify\naivebayes.py in classify
(self, featureset)
            def classify(self, featureset)
     88
                  return self.prob_classify(featureset).max()
  -> 89
     91
            def prob classify(self. featureset):
d:\users\owner\appdata\local\programs\python\python39\lib\site-packages\nltk\classify\naivebayes.py in prob clas
sify(self, featureset)
                featureset = featureset.copy()
     95
                for fname in list(featureset.keys()):
  -> 97
                       for label in self._labels:
   if (label, fname) in self._feature_probdist:
     99
                             break
```

KeyboardInterrupt:

```
In [16]: print(len(NOTpositive))
    print(NOTpositive[:10])

print(len(NOTnegative))
    print(NOTnegative[:10])
```

8658

['perfect for new parents.', 'this book is such a life saver.', 'see other things that are must haves for baby at [...]\nhelps me know exactly how my babies day has gone with my mother in law watching him while i go to work.', 'i bought this a few times for my older son and have bought it again for my newborn.', 'excellent for dads to keep track as my husband can quickly forget what time he fed our son.', "it's also nice to be able to look back on previous days and weeks for eating and sleeping pattern s.', 'a lot times i need struggle trying to read what the caretaker wrote in because the spaces go together.', 'my 3 month old son spend half of his days with my mother and half with a neighbor while i worked.', 'this book is perfect!', 'the baby tracker brand books are the absolute best of the trackers that are available.']

["we were able to keep track of baby's feeding, sleep and diaper change schedule for the first two and a half months of her li fe.", 'made life easier when the doctor would ask questions about habits because we had it all right there!', 'it has been so helpful to be able to go back to track trends, answer pediatrician questions, or communicate with each other when you are up a t different times of the night with a newborn.', 'i think it is one of those things that everyone should be required to have before they leave the hospital.', 'we went through all the pages of the newborn version, then moved to the infant version, and will finish up the second infant book (third total) right as our baby truns!.', 'it also has a section for her to write notes and let me know anything she may need.', "i couldn't be happier with this book.", 'this is super easy to use and helps me keep track of his daily routine.', 'when he started going to the sitter when i went back to work, it helped me know how his day went to better prepare me for how the evening would most likely go.', 'when he was sick, it help me keep track of how many diaper s a day he was producing to make sure he was getting dehydrated.']

Results and Analysis

The SL feature has produced an accuracy of 0.764 and the negation word feature has an accuracy of 0.78. In the classification of positivity, the SL feature classified ~27k positives and ~56k negatives whereas negation word feature classified ~8.6k positives and ~22.4k negatives. For this analysis, I will just use the NOTnegatives.txt and NOTpositives.txt produced from the negation feature.

Below are the reviews classified as negative. We can see some sentences are indeed negative. For example, "problem was i could not find the adhesive to attach the thing to the table." This is most likely due to the word 'problem' and 'could not.' However, we can also see some positive sentences, such as "good product." Many of the sentences are classified as negative because of the negation words. We can find this easily in the reviews.

```
i game it only four stars because the lites arrived with toom of it missing, that was however resolved, the product does what it is supposed to.

We can struck like the copy claer one i tried before. Sticks to my wood coffee table and glass trie on ty stand. Answer't tried to take it off yet so i don't know what that's going to be like but it was def worth it.

We have a structure of the absolute or with the but it was def worth it.

We have a structure of the star of the star of the but it was the thing to the table.

We then attached the corner pieces to the cut out parts using gerilla gles and then clean packing tape to re-enforce the hold.

We have a straight on the table, it has part usengy give to not not and off but not endough the lo will poil lit off easily and happ plus for us...no damage to our pottery burn table!

Once ig to se phones in all tatch but for the moment it is serving its purpose and we are very tappy!

A structure the plus of the table, it has part usengy give to not one of other plus of the products.

We allow the price landsare products.

We allow this one done of an off-reside to cover on the star page land to the country of the price lands are country to the star page lands to inches left or one or the star page land to the country of the price lands are country to the star page lands to inches left or one or the star page land to the country of the price lands are country to the star page lands to inches left or one or the star page land to the country of the price lands are country to the star page lands to the country of the star page lands to the country of the price lands are country to the star page lands to the country of the page lands to the page lan
```

On the other hand, the positive reviews are mostly actually positive reviews. We hardly find any contradictions. One noticeable thing we can find is that many neutral comments are classified into negative reviews. This is a main reason why the difference between the number of positives and negatives are huge. Another reason may be that such reviews are for baby products, but the data set we used was for movie reviews. Moreover, it also depend on the training features we used. For example, the highest accuracy we got was

78%, so if we use a better set, we may be able to get a better result.

```
/8%0, SO II We use a Detter set, we may be able to get a Detter result.

around the outside.

we use it daily.

this sterilizer is absolutely wonderful, especially if you take your baby on the road, i.e.
plus, unlike a bottle warmer appliance, there are no moving part or electrical components to wear out or break, so we expect we'll be using ours for years/babies to come.

this is such a great item and i can't say enough good things about it!

this is signest for anyone with the avent bottles.

after a couple weeks of use she decided to get this sterilizer.

it is fast and simple to use and holds a lot of bottles as well as my breast pump accessories.

it fits all of her bottles and pacifiers in one process and it fits perfectly in my small kitchen when i am not using it.

you will get good use out of it.

love the convience and ease of use.

very convient and can do 6 bottles at a time or even baby spoons, forks & even bowls or cups your choice.

i am not a cleam freak who sterilizes everything my son touches.

after a brief, and horrible stint with stovetop sterilizing, i began to use the dishwasher to clean my son�s bottles.

there is no way to confuse the stacking order: the base, the bottle rack, and the lid.

keep the manual handy, as the loading order for large bottles, differs from the 4-ounce size.3.)

in 4 or 5-minutes it�s done.
keep the manual handy, as the loading order for large bottles, differs from the 4-bunce size.,

1A 40 To -indinues it&B65533;s done.

1AB65533;s hot!

plus, any steam that&B65533;s emitted can burn.

once the cooling time has passed, 2-minutes or so, remove the lid carefully.

if it&B65533;s top-drawer dishwasher safe, than it&B65533;s probably sterilizer safe.

some might complain about the price of the sterilizer, but considering two, 9-ounce bottles come with the unit, it may seem more reasonable knowing they retail for $4.00 a piece.

the sterilizer works great, simply, and fast!

the express steam sterilizer is really the easiest and most efficient way to go if you want to be assured that your baby's bottles are truly clean.

when i first started using bottles i would use the sterilizer between dishwasher uses.

it is very quick and completely uncomplicated.

the avent bottles fit perfectly.

I now use the dr. brown's bottles and just have to lay them on their sides.

I sterilized bottles and breast pumps the old fashioned way with my first child and in the begining w/ my 2nd child.

this is a very simple, fast, effective and economical way to get just about everything for baby sterilized.

this is the most wonderful item i have gotten.

since i express milk and freeze it, it is reassuring knowing that the storage containers are sterile.

I also love it because you never know if you are washing the bottles and pacifiers and such with warm enough water (i am a first time mom).

It is so easy to use and comes with extra bottles.

It can sterilize 5 small bottles & injules 5 minutes at a time!!

I am happy to say that neither of us has suffered from thrush or any infection.it's also wonderfull y, too.i would highly recommend.

the wattage is simple to use, just look inside your microwave on the bottom with all the info and it tells you what wattage it is and them refer to the manual to see the time.

highly recommended
  the wattage is sample to use, just look insize your airrowave on the bottom with all the into and it tells you what wattage it is and them refer to the manual highly recomended
we initially bought this sterilizer for use with our first child born in 2005.
i loved it then, but i'm even more impressed with some updated details in the version we just bought for our second child.

now, easy-to-use latches have made doing both completely hassle-free.

that the sterilizer comes with bottles is pure bonus.
i bought this sterilizer when i was using avent bottles and it worked great.

our microwave is 1200 watts, so using this sterilizer saved me a ton of time...bottles were done in 3 minutes!
it saves my skin from hand washing in scorching hot water (but i still give things a quick rince in warm soapy water before tossing them in the sterilizer).
it is a lifesaver when pumping and bottle feeding the only drawback is the space that it takes up - it is about 12 inches in diameter and 6 inhes tall.
this is a great item if you don't run the dishwasher alot.

we use this for home and travel.
good item.
finally, and perhaps most important of all: this is an essential piece of baby gear.
this is a great thing to have and we still use it a few times a week and my son is 10 wks old.
```