STC LAB
Sentiment Analysis using Python

Pratyush Kaware
TE EXTC
2017120029

Vedant Kayande
TE EXTC
2017120030

Pranit Mahajan
TE EXTC
2017120037

***Abstract*—To find if a review is positive or negative. The phrases in short movie reviews convey different sentiments. For example, a phrase denotes positive sentiment about the film while another one treats the movie as not so great (negative sentiment). Sometimes there's no such word in that phrase which can tell you about anything regarding the sentiment conveyed by it. Hence, that is an example of neutral sentiment. Our aim is to classify according to this.**

I. Introduction:

he The promise of machine learning has shown many stunning results in a wide variety of fields. Natural language processing is no exception of it, and it is one of those fields where machine learning has been able to show general artificial intelligence (not entirely but at least partially) achieving some brilliant results for genuinely complicated tasks. Until recently, sentiment analysis was a niche technology only accessible to technical people with coding skills and background in machine learning. This is no longer the case thanks to the rise of a variety of tools that can be leveraged to get the data and run sentiment analysis models. Essentially, sentiment analysis or sentiment classification fall into the broad category of text classification tasks where you are supplied with a phrase, or a list of phrases and your classifier is supposed to tell if the sentiment behind that is positive, negative or neutral. Sometimes, the third attribute is not taken to keep it a binary classification problem. The massive admiration and acceptance of social media tools and applications, new doors of opportunity have been opened for using data analytics in gaining meaningful insights from unstructured information.The application of sentiment analysis in the era of big data have been used a useful way in categorizing the opinion into different sentiment and in general evaluating the mood of the public.

II. today's problems

With an increase in social media, a lot of businesses depend and survive on it. To cater proper services to their consumers and to maintain customer satisfaction, a lot of people waste their time designing and analyzing surveys.

III. solution to the problem

Sentiment analysis eliminates the need of both designing and analyzing surveys and saves human hours.Scrutinizing every detail available on social platforms is a daunting task. That information comes from individuals that sample products, services, or those who view television commercials and take to social media to express their emotions. In most cases, the results at the end of such a process are not reliable. Therefore, to overcome this challenge, sentiment analysis gives you an overview of what individuals on social media think about your products, services, or plans.

IV. Working of the Code

A. Importing the Dataset

Using NLTK (Natural Language Tool Kit) to import moviereviews dataset. To import stop-words, we used the stopwords dataset.

B. Data Pre-processing

Filtering the data set to only include relevant features. This can be done by removing some words called fluff words.
Fluff words: A bag-of-words representation of a document does not only contain specific words but all the unique words in a document and their frequencies of occurrences. A bag is a mathematical set here, so by the definition of a set, the bag does not contain any duplicate words. Here is an example bag-of-words from the previous example. But for this application, you are only interested in the bold words as mentioned earlier, so the bag-of-words for this document will only contain these words. The words that you found out in the bag-of-words will now construct the feature set of your document. So, consider you a collection of many movie reviews (documents), and you have created bag-of-words representations for each one of them and preserved their labels (i.e., sentiments - +ve or -ve in this case). More recently, new feature extraction techniques have been applied based on word embeddings (also known as word vectors). This kind of representations makes it possible for words with similar meaning to have a similar representation, which can improve the performance of classifiers.

C. Training the model

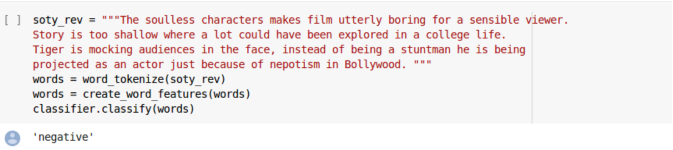
In machine learning, naive Bayes classifiers are a family of simple ”probabilistic classifiers” based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They are among the simplest Bayesian network models. The first step in a machine learning text classifier is to transform the text into a numerical representation, usually a vector. Usually, each component of the vector represents the frequency of a word or expression in a predefined dictionary (e.g. a lexicon of polarized words). This process is known as feature extraction or text vectorization and the classical approach has been bag-of-words or bag-of-ngrams with their frequency. We are going to use Naive Bayes Classifier to train the dataset Naive Bayes has two advantages:

- Reduced number of parameters.
- Linear time complexity as opposed to exponential time complexity.

The primary objective of the Bayes rule here, i.e. to find out the maximum posterior probability/estimate of a certain document belonging to a particular class. Think it this way - what is the probability of the occurrences of these words (features) given the class c. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

D. *Results:*

We are getting an accuracy of 71.99 percentage. The classifier determined these features as important.



```
[ ] soty_rev = """The soulless characters makes film utterly boring for a sensible viewer.
    Story is too shallow where a lot could have been explored in a college life.
    Tiger is mocking audiences in the face, instead of being a stuntman he is being
    projected as an actor just because of nepotism in Bollywood. """
    words = word_tokenize(soty_rev)
    words = create_word_features(words)
    classifier.classify(words)
```

```
'negative'
```

Fig. 1. Features used

V. Conclusion

The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments. This is usually measured by variant measures based on precision and recall over the two target categories of negative and positive texts. However, according to research human raters typically only agree about 80 percent of the time (see Inter-rater reliability). Thus, a program which achieves 70 percent accuracy in classifying sentiment is doing nearly as well as humans, even though such accuracy may not sound impressive. If a program were ”right” 100 percent of the time, humans would still disagree with it about 20 percent of the time, since they disagree that much about any answer. An estimated 80 percent of the world’s data is unorganized, much of that in textual form such as emails, support tickets, chats, social media, surveys, articles, and documents. Manually sorting through it all would be difficult, expensive, and impossibly time-consuming. Using sentiment analysis allows us to make sense of this chaos through automation, yielding actionable insights otherwise unattainable.

VI. Code

A. *Code link*

https://github.com/stclab-projects/stcproject-sentiment-analysis-of-movie-revies.git

VII. Future Scope

A. *Finding hot keywords:*

Opinion mining can majorly help in discovering hot search keywords. This feature can help the brand in their SEO (Search Engine Optimization). This means that opinion mining will help them make strategies about, how their brand will come up among the top results, when a trending or hot keyword is searched in a search engine.

B. *Better services:*

Text mining can provide a filter about, which service of the company is getting more negative feedback. This will help the company to know, what are the problems arising with that particular service. And based on this information the company can rectify these problems.

C. *Get to know what’s trending:*

This will not only help the company to stay updated and connect more with the audience, but it will also facilitate the rise of new ideas, for developing new products. This will allow the company determine what the majority of the audience demands, and develop a product according to these demands.

VIII. References

- https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
- https://monkeylearn.com/sentiment-analysis/