Unsupervised Clustering For Sentiment Classification

Nghi Le, Nakul Gupta, John D'Ambrosio, Sanghun Kim, Gary Zhou

USC Viterbi School of Engineering
Department of Computer Science

Motivation

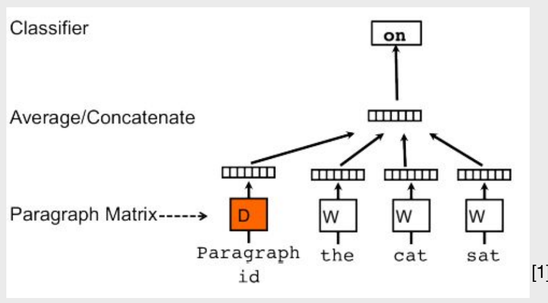
- As the amount of data grows massively, the burden of labeling data for companies and governmental organization has grown exponentially.
- Thus, unsupervised learning is becoming an important research direction as these techniques can learn patterns in data without labels.
- If these methods' performance are acceptably effective in tasks such as classification, then the cost to acquire data labels for organizations can be extremely low.
- This leads to better access to the power of AI/machine learning for everyone.
- Our project aims to study whether unsupervised learning can be used for classification tasks, when a popular language model vector embedding is used as a text representation

Setup

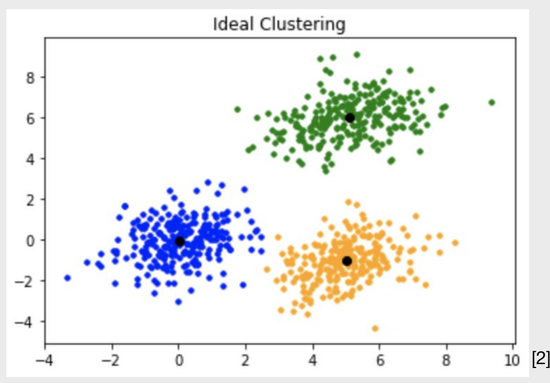
- We used the IMDb movie review dataset to conduct our unsupervised sentiment analysis. We chose this dataset because it is often used as a benchmark for sentiment analysis models.
- The data is high polar, meaning it is very easy for humans to distinguish between positive and negative sentiment movie reviews.
- The dataset contains 50,000 training and testing samples with binary sentiment labels.
- We used Doc2Vec to encode each movie review into a vector. It is a method to generate numeric representations of text documents regardless of the length of the text documents.
- In Word2Vec words of similar meaning will be closer in distance in the vector space, and the word vectors represent the machine learned concepts of a word. In Doc2Vec, the document vector represents the concept of a document.
- Doc2Vec works just like Word2Vec. Word2Vec uses the CBOW and Skip-Gram models to train a word matrix that represents the learned concepts of words. The Doc2Vec model trains an additional document matrix to represent documents numerically.
- The document vector captures the learned concepts of a document. We were hoping that Doc2Vec can generate document vectors that form clear cluster groups to help us identify positive and negative sentiments.

Models & Methods

- An Doc2Vec embedding model was trained using the first 10,000 rows of the shuffled IMDB sentiment dataset. Then, we put all 50,000 rows of the dataset into the Doc2Vec model to create the vectors.
- The Doc2Vec model was trained on the following hyperparameters:
  - learning rate = 0.025
  - minimum learning rate = 0.00025
  - feature vector size = 500
  - max epochs = 10.
- We selected K-Means clustering algorithm for our unsupervised clustering method.
- The Doc2Vec vectors were clustered using the K-Means algorithm with two clusters, as the dataset had only two sentiment labels (positive and negative).
- These assigned clusters were compared against the document vectors' true sentiment labels to find a measure of clustering accuracy.
- We used the Monte Carlo simulation method to run 50 trials to measure the average performance of the K-means clustering algorithm on predicting the correct sentiment labels.



[1]



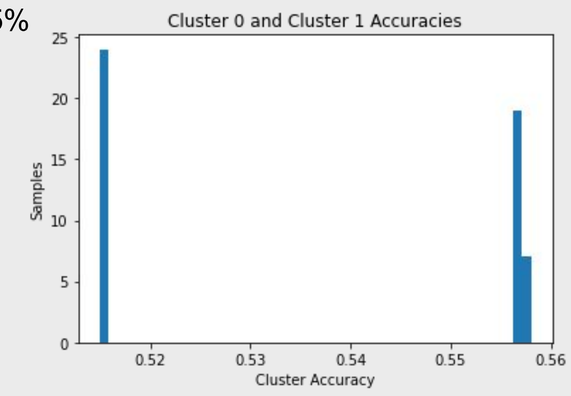
[2]

Results

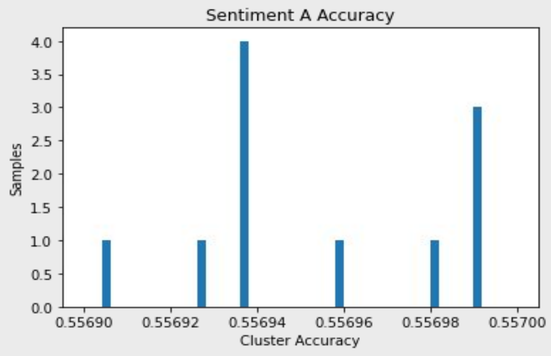
- A Monte Carlo analysis was performed with 50 trials, with each trial fitting the entire dataset.
- Each trial returned a percentage that corresponds to the majority sentiment for each cluster. Each cluster then represents that sentiment.
- The percentages act as an accuracy score for each cluster to predict a given sentiment.
- After 50 trials, the mean and standard deviation are calculated for each cluster to account for clusters changing from trial to trial

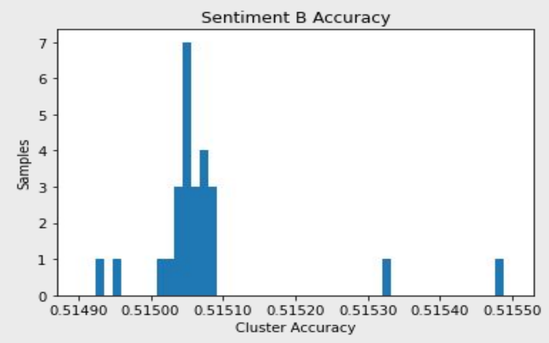
Cluster	Mean	Standard Deviation
0	.535	.021
1	.536	.021

- The results show that on average, given either sentiment, a cluster has a 54% accuracy of prediction
- Looking at the entirety of accuracy scores from each trial and plotting on a histogram, we can see that there are two distributions centered around 51% and 55%



- Examining each distribution more closely, we can see somewhat of a gaussian distribution





Conclusion & Future Work

- These results indicate that with Doc2Vec method of text embedding, document vectors do not cluster naturally based on sentiment. Natural Doc2Vec embeddings learned from a standard language modeling task are not particularly useful for the task of unsupervised sentiment classification.
- Future steps:
  - Different document embedding methods instead of Doc2Vec,
    - like bag of words, bag of ngrams, TFIDF
  - More unsupervised learning methods
    - like Gaussian Mixture Models, DBSCAN, or Spectral Clustering
  - Other categories of Text Classification
    - Truthful versus Deceptive, Hate Speech, Sarcasm
- Sentiment embeddings are an area of active research and can also be studied with respect to the unsupervised sentiment classification task.

[1] https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e, [2] https://www.geeksforgeeks.org/ml-k-means-algorithm/