Google News Economy Dataset

Hamza Al-Habash

ID:19110034

HTU: School of Computing and Informatics

Date: 2021/12/17

Count-Vectorizer

- It transforms a given text into a vector, based on the frequency/count of each word that occurs in the entire text corpus.

- The Count-Vectorizer model can be used to convert each word in each text into vectors in case we have multiple texts corpus.

Term frequency–Inverse Document Frequency (TF-IDF)

- It’s a Vector Space Model(VSM) which represents text units as vectors of numbers.

- The TF-IDF approach quantifies words in a set of documents. and it's used to signify each word's importance in the document and corpus.

Word Embedding [Word2Vec]

- Word Embedding is a representation of document vocabulary which can capture the context of a word in a document, semantic similarity, relation with other words.

- The Word2vec model groups the words with similar meanings together, and words with different meanings far from each other, by representing the meaning of the words as real-valued vectors in a vector space.

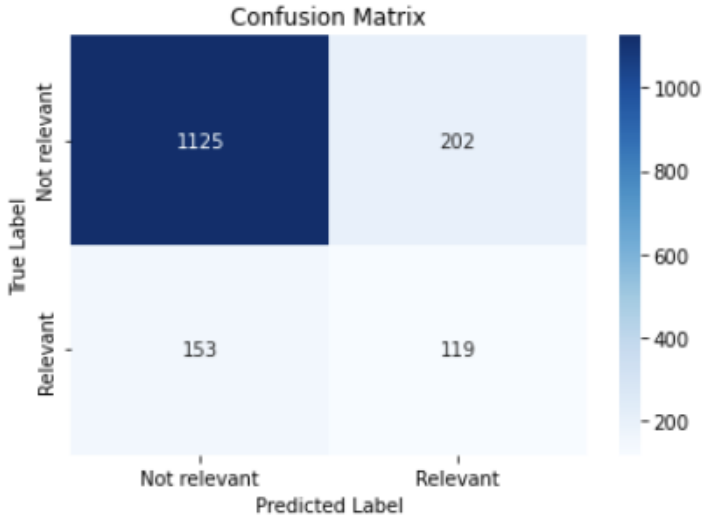
Main Performance Conclusions

- I've used the Naive bayes classifier in classifying the economy news articles with all the models.

- I found that the Count-Vectorizer model accuracy performance was [0.7779862414008756] with a confusion matrix as represented below.
- On the other hand, I found that the TF-IDF & Word Embedding W2V models surprisingly have the same accuracy performance which was [0.8298936835522202] and the same confusion matrix for each is represented below.

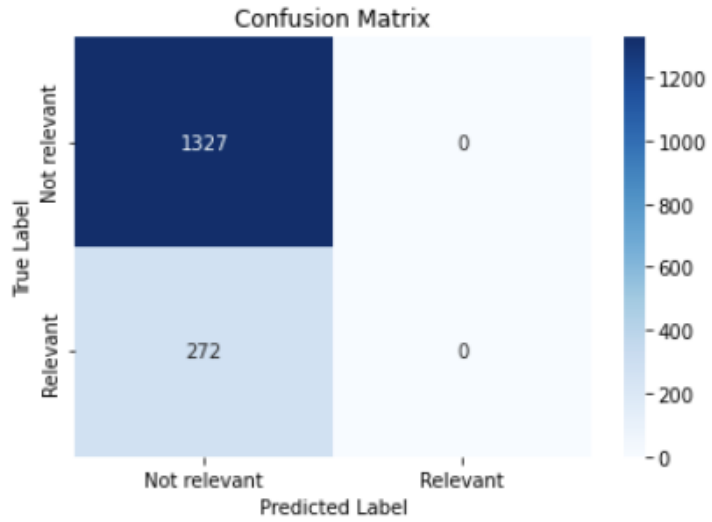
- **As a result**, I can conclude that the TF-IDF and Word2Vec models both were better than the count-vectorizer model with a higher accuracy and with a better matrix, since both had a more true and less false overall result. Thus, they would be the best models to use to analyze our data.

Accuracy: 0.7779862414008756



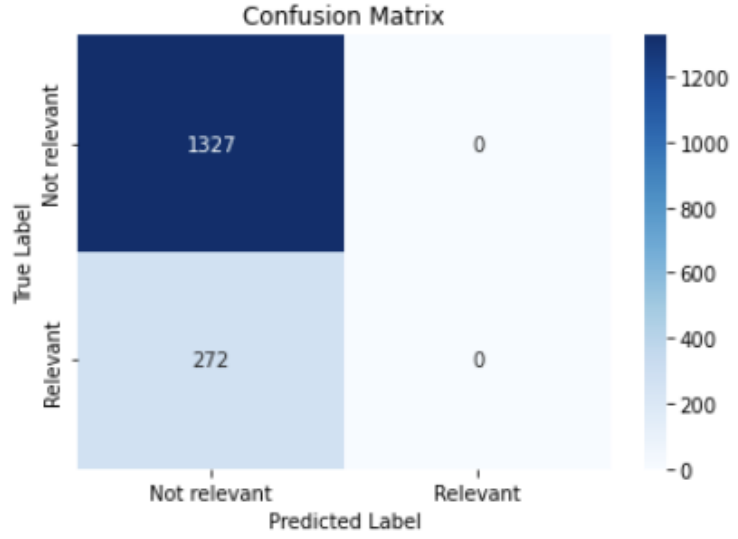
Count-Vectorizer

Accuracy: 0.8298936835522202



TF-IDF

Accuracy: 0.8298936835522202



Word2Vec