Opinion-Fact Classification

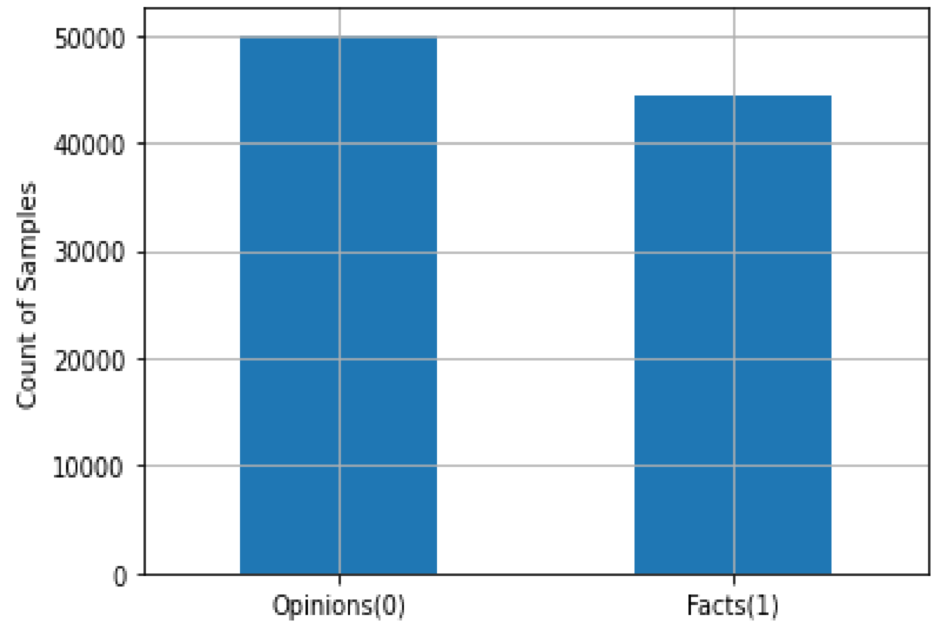
K. Mani Kumar Reddy (MT19065), K. Sarath Chandra Reddy (MT19037), K. Murali Krishna (MT19132)

Problem Statement

- In the present-day technology huge amount of data is being generated every day. So, it's turning out to be a challenging task to handle text-based data.
- In the world of text-based sentences it is not that simple to differentiate between fact and opinions.
- So, our project is to build the model that classifies/identifies facts from/and opinions in the given text by using various machine learning and deep learning techniques.

Dataset Description

- The dataset we will be using for this project is hand annotated. We considered the data from "movies" domain and annotated them into opinions and facts. Here, the plot of a movie is considered as fact. whereas the review of an individual for a movie is considered as opinion. https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv
- The dataset contains 94,379 samples which are facts or opinions.
- Dataset has opinion count of 50,000 whereas facts of 44,379.
- The dataset has train, cross-validation & test splits



Preprocessing Techniques

- Stop-Word removal
- Case Conversion
- Tokenization, lemmatization
- Removal of alpha-numeric words and special characters.
- Removal of words of length less than 3.

Learning Techniques

- K-NN (BOW & TFIDF) - Baseline
- Naive Bayes (BOW & TFIDF)
- Decision Trees (BOW & TFIDF)
- SVMs (BOW & TFIDF)
- LSTM (Long Short-Term Memory)
- Deployment of best model using flask

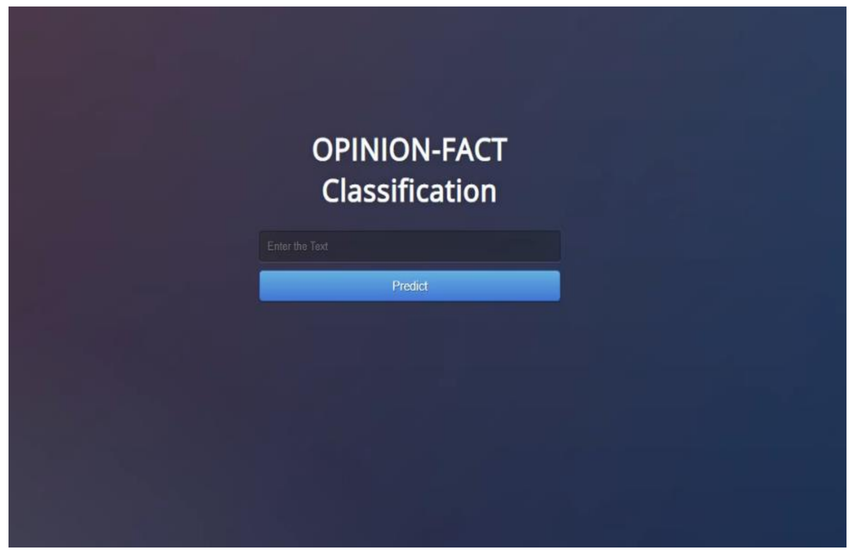
Evaluation Metrics Used

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion matrix
- Binary-Cross Entropy loss (for LSTM)

Results

Model Implemented	Word-Embedding Used	Precision achieved on Test Data	Recall achieved on Test Data	F-Score achieved on Test Data	Accuracy achieved on test data
K-NN (baseline)	TF-IDF	0.6387	0.506	0.351	50.6%
K-NN	BOW	0.832	0.754	0.739	75.45%
Naïve Bayes	BOW	0.814	0.795	0.792	79.50%
Naïve Bayes	TF-IDF	0.811	0.788	0.7844	78.85%
Decision Trees	BOW	0.9062	0.9050	0.9046	90.46%
Decision Trees	TF-IDF	0.9192	0.9180	0.9176	91.76%
SVM	BOW	0.9576	0.956	0.9569	95.7%
SVM	TF-IDF	0.9542	0.953	0.9539	95.4%
LSTM	Rank of word in the vocabulary	0.9866	0.9870	0.9867	98.62%

Deployment


OPINION-FACT
Classification

- We have deployed the LSTM model (for its better performance compared to other models) using flask web frame work. The created web-page can be seen in the above picture.

References

- Most of the earlier research on opinion classification i done by Wiebe and his collegues (Weibe et al., 1999). they proposed methods for discriminating subjective and objective features.
- Hatzivassiloglou and McKeown proposed an un supervised model for learning positively and oriented adjectives with accuracy over 90%.
- A similar study was conducted by Ahmet Aker et in his paper titled "Beyond opinion classification: extracting facts and opinions from health forums".
- https://www.youtube.com/watch?v=UbCWoMf80PY&t=692s