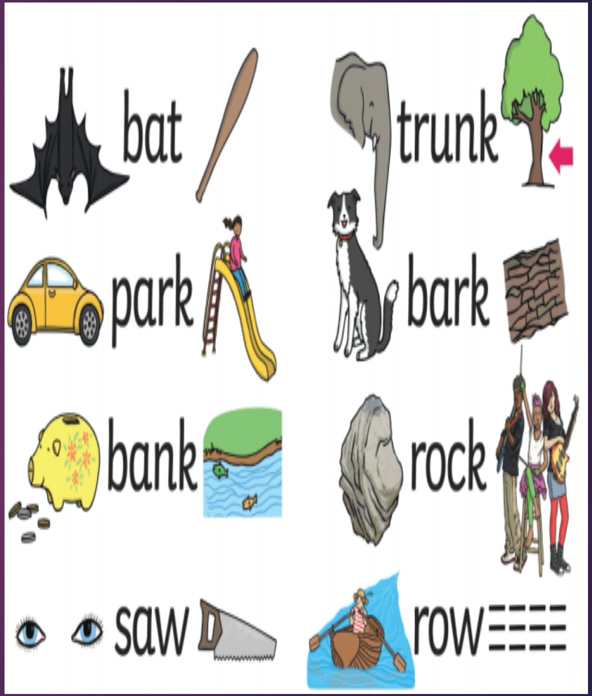
Homonyms Problem In NLP



What is Homonyms problem??

- The **homonyms problem** in NLP is the challenge of correctly interpreting words with the same spelling or pronunciation but different meanings, requiring context to resolve ambiguity (e.g., "bat" as an animal vs. sports equipment).

- Examples :

➢ "You hate anything that hurts your loved ones." (POSITIVE)
➢ "They hate seeing others get hurt." (POSITIVE)
➢ "I don't love that you work so hard, but your dedication is admirable." (POSITIVE)
➢ “I hate the selfishness in you” (NEGATIVE)
➢ "I hate any one who can hurt you ” (POSITIVE)

Data Set

- **stanfordnlp/sst2**
  - consist of 11,855 single sentences extracted from movie reviews,annotated by 3 human judges and the complete records on it 67348

- **Custom Data**
  - custom data was generated to specifically target challenging homonym examples. This custom dataset comprises 1,185 sentences created using ChatGPT.

Examples:

- "You hate anything that hurts your loved ones." (POSITIVE)
- "They hate seeing others get hurt." (POSITIVE)
- "I can't tolerate your constant complaining." (NEGATIVE)

Data Set

- **Training Set Size:** 68534 sentences (SST2+Custom data)
- **Validation Set Size:** 872 sentences.
- **Test Set Size:** 20 sentences contains homonyms examples

Data Preprocessing

- **Lowercasing**: All text was converted to lowercase to ensure uniformity and reduce the complexity of the text.
- **Removing Punctuation and Numbers**: All punctuation marks and numbers were removed from the text to reduce noise and focus on the meaningful content of each sentence.
- **Removing Stop Words**: Common stop words (e.g., "the," "is," "in") were removed to focus on the significant words that contribute to the sentiment of the text.

Data Preprocessing(BI-LSTM)

- **Vocabulary Creation**: A vocabulary was created from the training data, which included a special <pad> token for padding.
- **Tokenization**: Each sentence in the dataset was tokenized using the created vocabulary, converting words into corresponding numerical tokens.
- **MaxLength:** Use tokenizer for max length computation to pad sentences Length 28
- **Padding:** All sentences were padded with the <pad> token to ensure that they have a uniform length. The <pad> token was assigned an ID of 0 to distinguish it from other tokens.

Data Preprocessing(BERT)

- **Tokenization**: The text data was tokenized using the DistilBERT tokenizer, which is capable of handling subword tokenization. This helps manage out-of-vocabulary words and maintain contextual understanding.
- **Maximum Length Calculation**: The maximum length of the tokenized sentences was computed to be 42 words. This ensures that the model can handle the longest sentence in the dataset while maintaining efficiency.

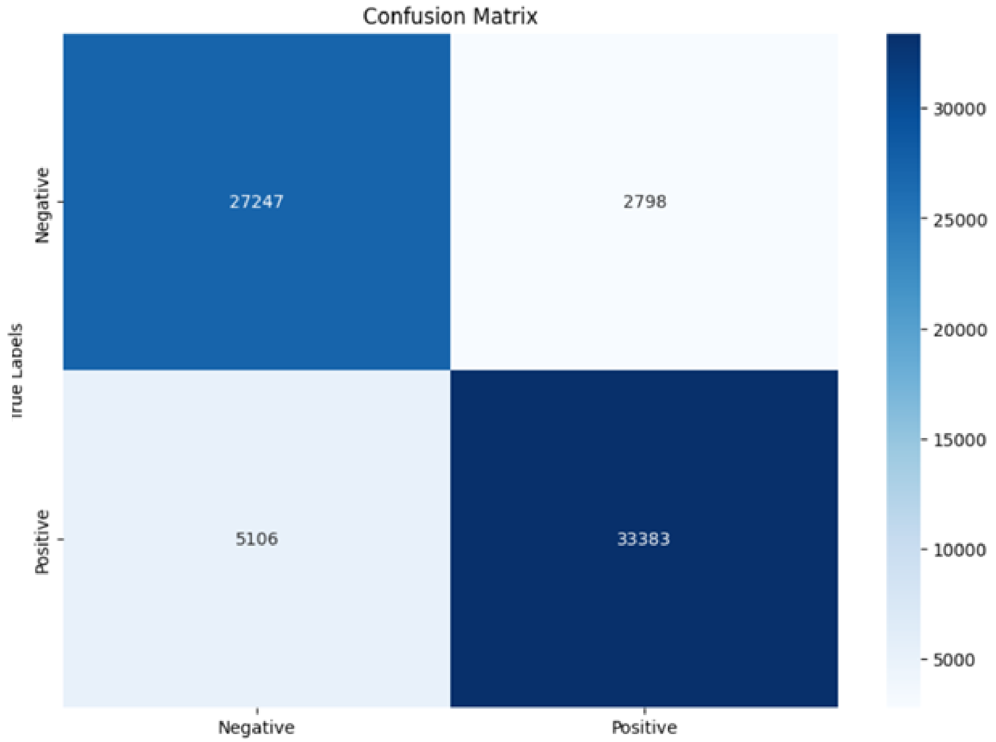
Training Details

- **Gradient Accumulation**: To manage memory efficiently and stabilize training, gradient accumulation was employed.
- **Mixed Precision:** The training process was optimized using mixed precision with a scaler, maximizing precision and computational efficiency.

Results(BI-LSTM-Training)

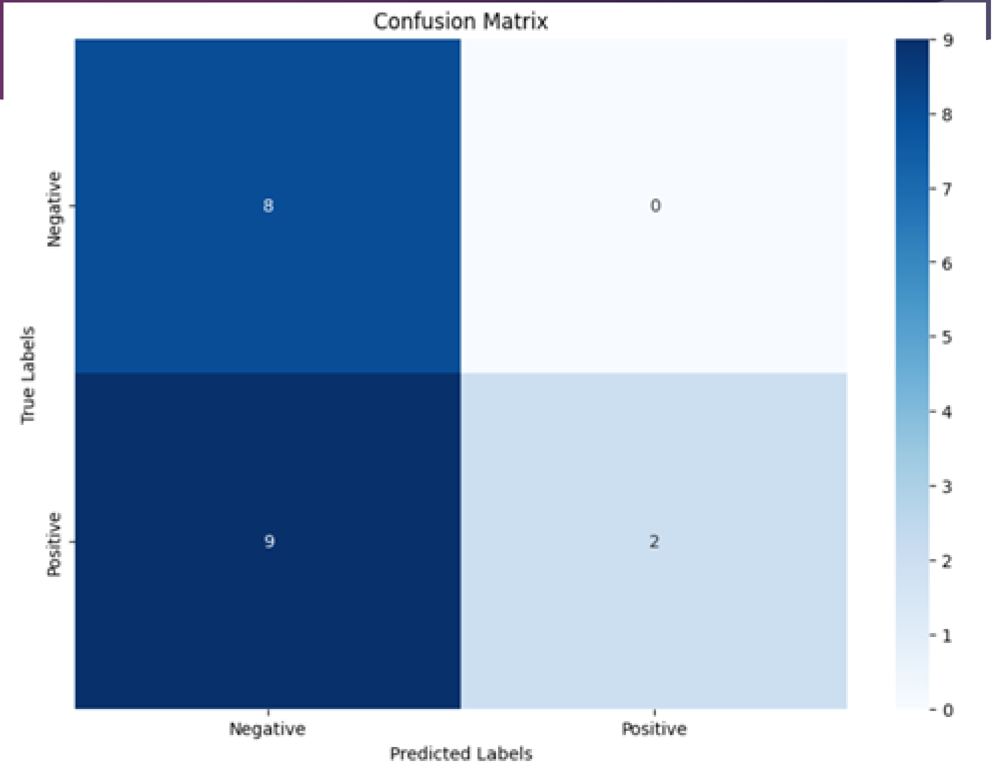
- Training F1 Score: 0.885
- Validation F1 Score: 0.785
- Training Confusion Matrix:

	precision	recall	f1-score	support
Negative	0.84	0.91	0.87	30045
Positive	0.92	0.87	0.89	38489
accuracy			0.88	68534
macro avg	0.88	0.89	0.88	68534
weighted avg	0.89	0.88	0.89	68534



Results(BI-LSTM-Testing)

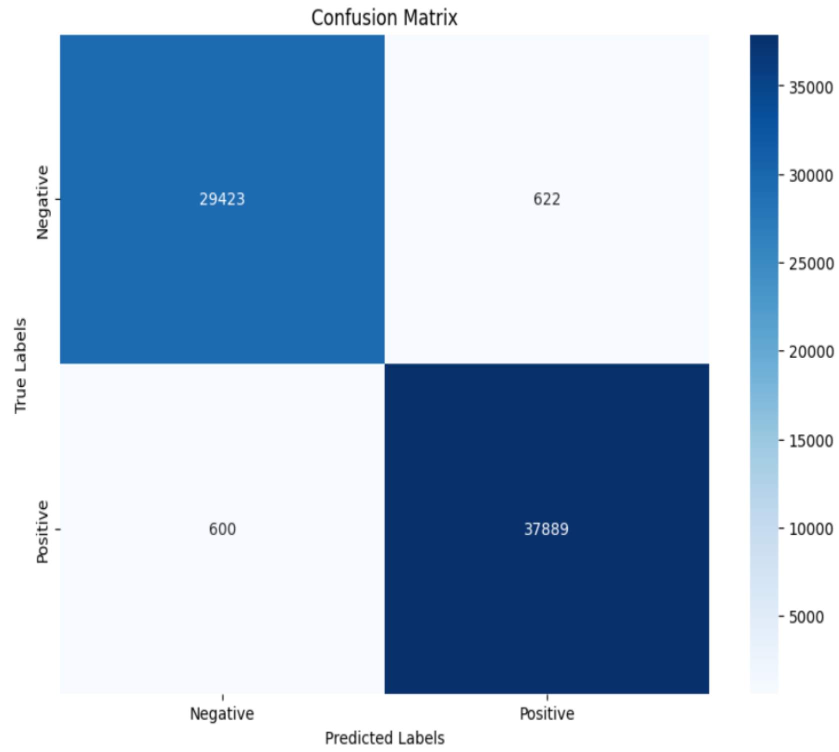
- **Testing F1 Score**: 0.4476
- **Testing Confusion Matrix**:



Results(BERT-Training)

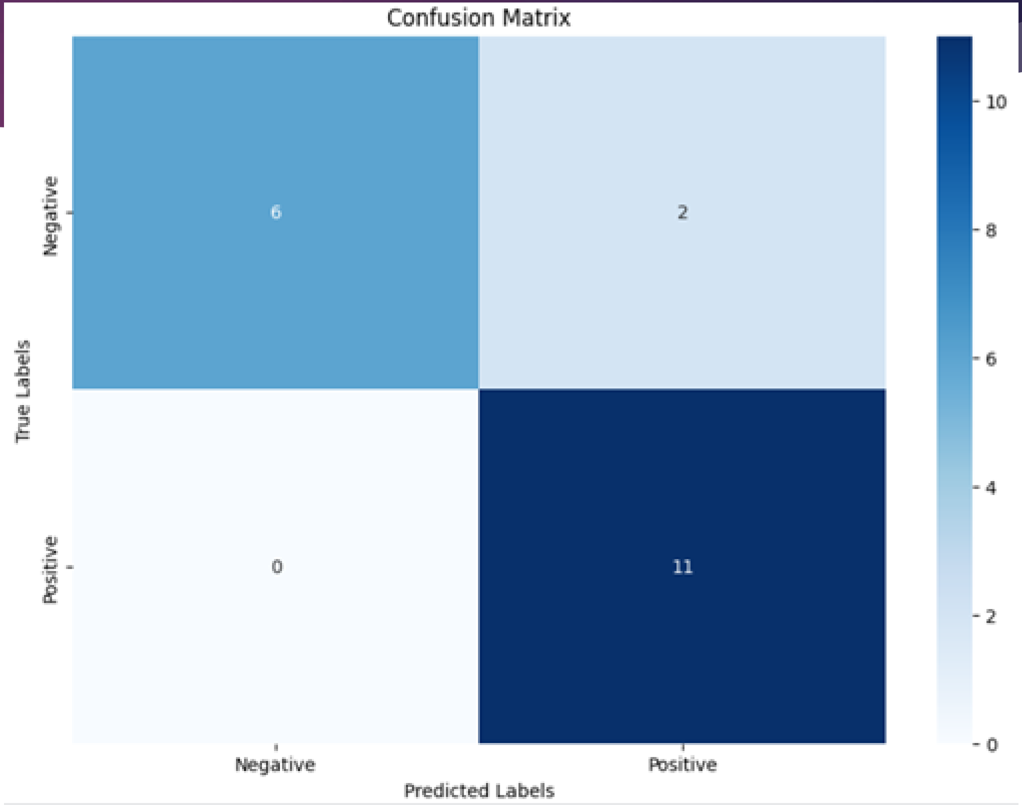
- **Training F1 Score**: 0.965
- **Validation F1 Score**: 0.843
- **Training confusion matrix**:

Negative	0.98	0.98	0.98	30045
Positive	0.98	0.98	0.98	38489
accuracy			0.98	68534
macro avg	0.98	0.98	0.98	68534
weighted avg	0.98	0.98	0.98	68534



Results(BERT-Testing)

- **Testing F1 Score**: 0.891
- **Testing Confusion Matrix**:



Results

	text	label	BILSTM	BERT
0	I love you	POSITIVE	POSITIVE	POSITIVE
1	I hate you	NEGATIVE	NEGATIVE	NEGATIVE
2	I hate the selfishness in you	NEGATIVE	NEGATIVE	NEGATIVE
3	I hate anyone hurt you	POSITIVE	NEGATIVE	POSITIVE
4	I hate anyone hurt you, you are my partner	POSITIVE	NEGATIVE	POSITIVE
5	I hate anyone hurt you, you are my love	POSITIVE	NEGATIVE	POSITIVE
6	I like rude people	NEGATIVE	NEGATIVE	NEGATIVE
7	I don't like rude people	POSITIVE	NEGATIVE	POSITIVE
8	I hate polite people	NEGATIVE	NEGATIVE	NEGATIVE
9	I don't hate polite people	POSITIVE	NEGATIVE	POSITIVE
10	I love when you are honest	POSITIVE	POSITIVE	POSITIVE
11	I hate when you are honest	NEGATIVE	NEGATIVE	POSITIVE
12	I don't hate when you are honest	POSITIVE	NEGATIVE	POSITIVE
13	I like how you always tell the truth	POSITIVE	NEGATIVE	POSITIVE
14	I hate how you always tell the truth	NEGATIVE	NEGATIVE	NEGATIVE
15	I don't like how you always lie	NEGATIVE	NEGATIVE	POSITIVE
16	I like how you never lie	POSITIVE	NEGATIVE	POSITIVE
17	I hate people who are kind	NEGATIVE	NEGATIVE	NEGATIVE
18	I don't hate people who are kind	POSITIVE	NEGATIVE	POSITIVE

Conclusion

- **Enhanced Context Sensitivity**: BERT effectively understands nuanced contexts, correctly classifying complex sentences (e.g., “I hate anyone hurt you” as positive), unlike BiLSTM.
- **Effective Handling of Long Sequences**: BERT excels with long sentences, accurately interpreting sentiments even when critical context appears late (e.g., “I hate anyone hurt you, you are my partner”).
- **Improved Negation Handling**: BERT’s contextual embeddings accurately identify sentiments in negated sentences (e.g., “I don’t like rude people”), while BiLSTM may misclassify them.
- **Context Embeddings vs. LSTM Embeddings**: BERT uses bidirectional transformers for richer context embeddings, capturing full word context and long-range dependencies better than sequential LSTM embeddings.

Resources

- Stanford Sentiment Treebank (SST-2) (Dataset)
- Contextual Embeddings
- BERT
- Attention Is All You Need