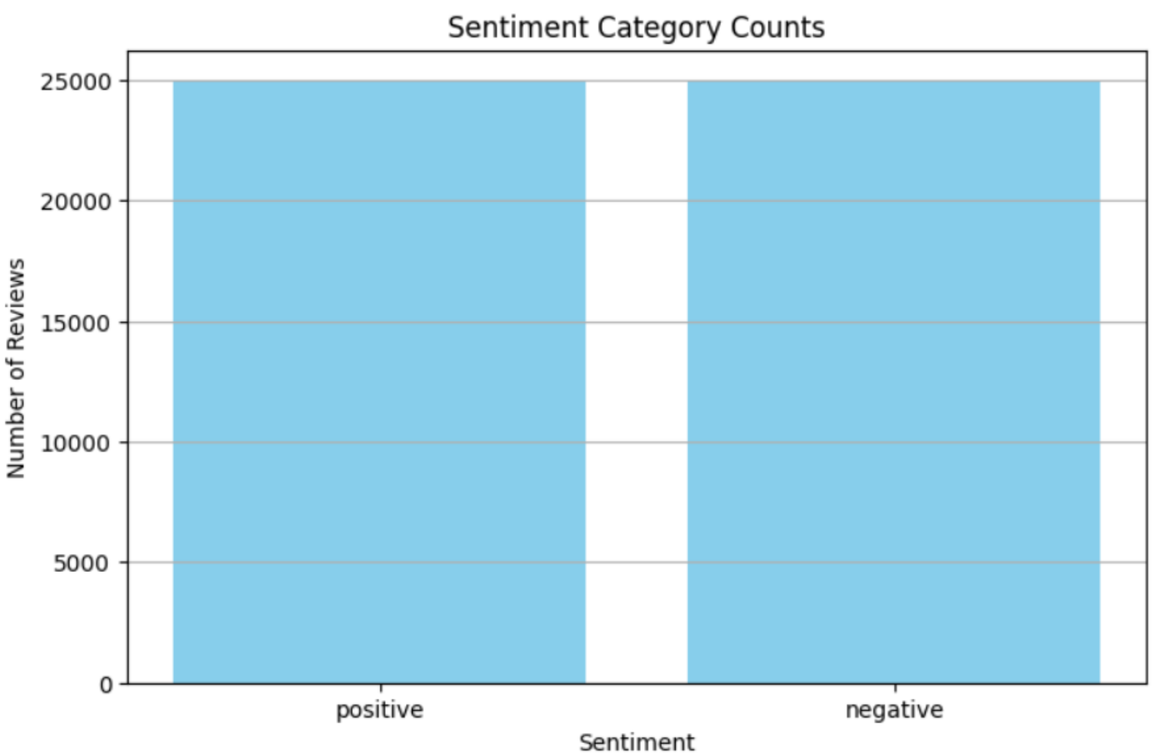
Load and Explore the Datasets

Dataset A: IMDB (50K Movie Reviews)

IMDB dataset having 50K movie reviews for natural language processing or Text analytics. This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing. So, predict the number of positive and negative reviews using either classification or deep learning algorithms.

- **Labels**: positive or negative (textual)
- **Text**: Full movie reviews
- **Format**: CSV (2 columns – review, sentiment)



Preprocessing

- Lowercase
- Remove URLs
- Remove mentions
- Remove hashtags
- punctuation/numbers
- tokenize
- remove stopwords.words('english')
  by nltk

```
def preprocess_text(text):
    text = text.lower()                       # Lowercase
    text = re.sub(r"http\S+|www.\S+", '', text)  # Remove URLs
    text = re.sub(r"@\w+", '', text)           # Remove mentions
    text = re.sub(r"#\w+", '', text)           # Remove hashtags
    text = re.sub(r"[^a-z\s]", '', text)        # Remove punctuation/numbers
    tokens = word_tokenize(text)
    tokens = [w for w in tokens if w not in stopwords.words('english') and len(w) > 2]
    return tokens
```

Research Opportunities/Tasks

a. Sentiment Classification

- **Binary classification**: Both datasets are labeled as **positive (1)** and **negative (0)**. You can train models like Logistic Regression, LSTM, or BERT.
- **Justification**: These datasets are large and well-labeled, suitable for supervised learning.

b. Domain Adaptation & Transfer Learning

- Train on **IMDB (movie reviews)** to study how well models transfer across domains.
- **Justification**: Real-world applications involve cross-domain texts (e.g., applying product review models to social media).

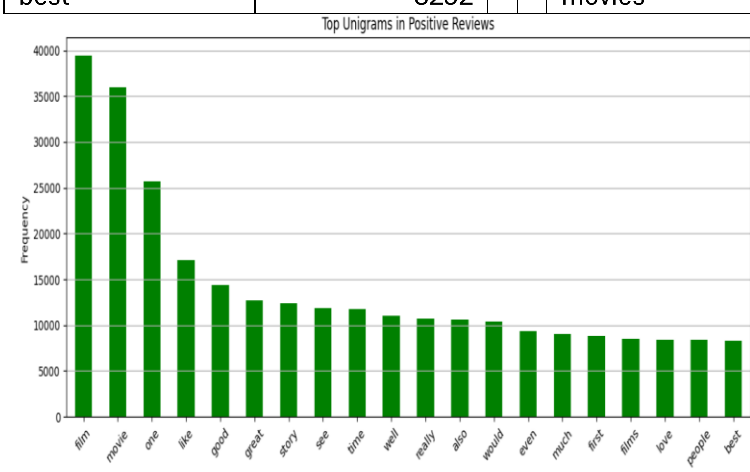
c. Preprocessing & NLP Pipeline Comparison

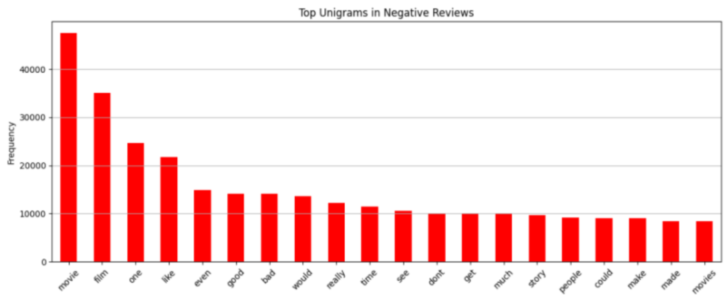
- Compare performance with different text-cleaning and preprocessing steps: stop word removal, stemming, lemmatization etc.

unique words list, Bigrams and Trigrams

Unigrams unique words top 20 in IMDB data set

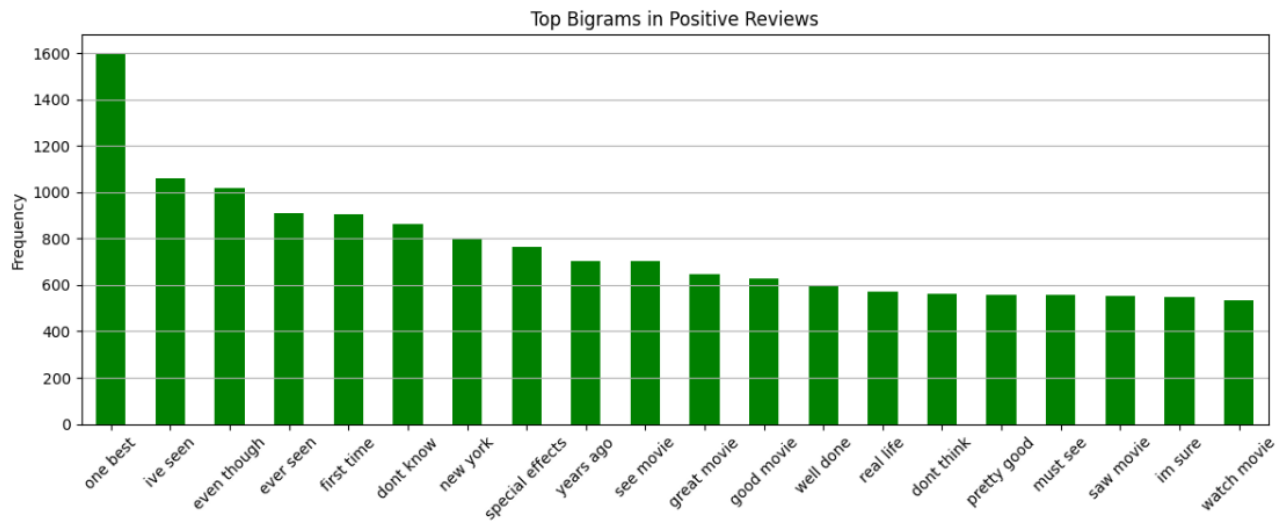
Top 20 Unigrams in Positive Reviews:				Top 20 Unigrams in Negative Reviews:	
word	freq			word	freq
film	39414			movie	47488
movie	36010			film	35042
one	25737			one	24636
like	17050			like	21771
good	14343			even	14918
great	12646			good	14141
story	12373			bad	14068
see	11868			would	13633
time	11777			really	12220
well	10979			time	11494
really	10676			see	10567
also	10545			dont	10029
would	10363			get	9996
even	9363			much	9898
much	8999			story	9672
first	8868			people	9115
films	8454			could	9031
love	8393			make	8985
people	8363			made	8391
best	8292			movies	8352

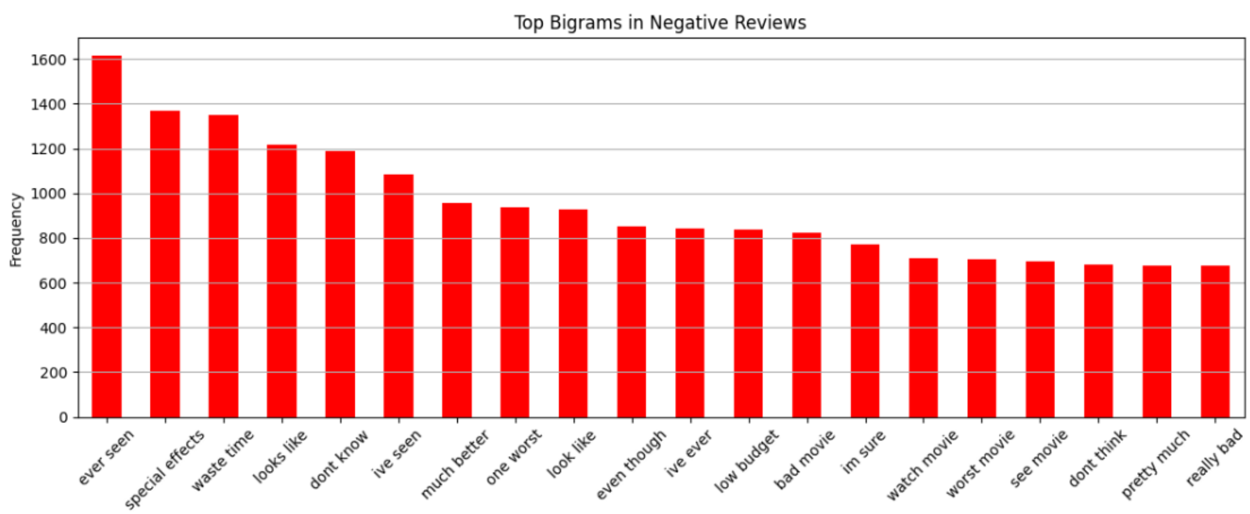




Binary Top 20 Bigrams IMDB DATA SET

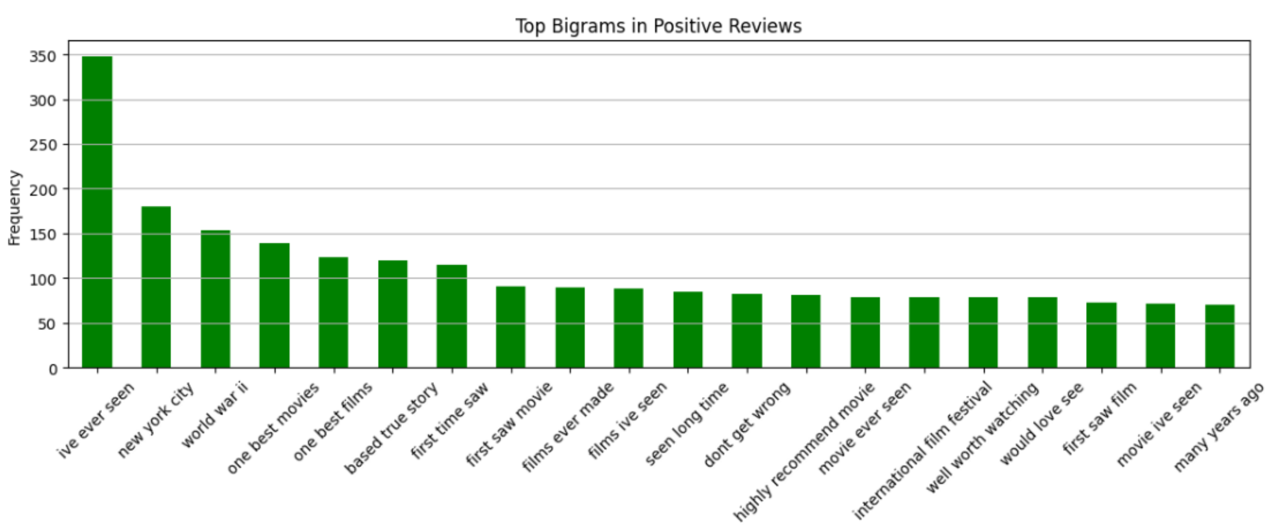
Top 20 Bigrams (Binary) in Positive Reviews:			Top 20 Bigrams (Binary) in Negative Reviews:		
word	freq			word	freq
one best	1599			ever seen	1614
ive seen	1061			special effects	1369
even though	1016			waste time	1349
ever seen	912			looks like	1219
first time	904			dont know	1191
dont know	862			ive seen	1086
new york	802			much better	955
special effects	763			one worst	936
years ago	703			look like	927
see movie	701			even though	852
great movie	647			ive ever	843
good movie	629			low budget	838
well done	601			bad movie	824
real life	571			im sure	771
dont think	563			watch movie	708
pretty good	556			worst movie	706
must see	556			see movie	694
saw movie	551			dont think	679
im sure	548			pretty much	677
watch movie	532			really bad	675

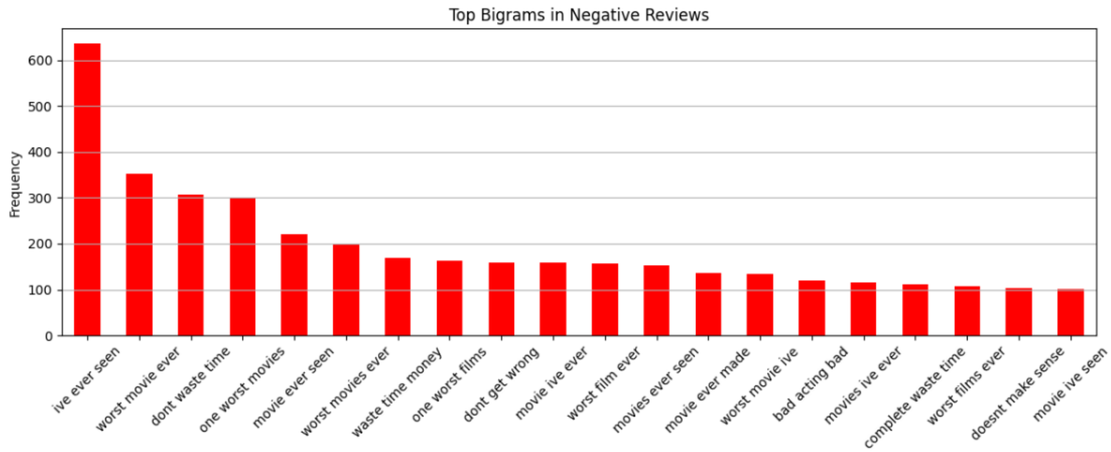

Top Bigrams in Positive Reviews
Frequency
0
200
400
600
800
1000
1200
1400
1600
one best
ive seen
even though
ever seen
first time
dont know
new york
special effects
years ago
see movie
great movie
good movie
well done
real life
dont think
pretty good
must see
saw movie
im sure
watch movie


Top Bigrams in Negative Reviews
Frequency
0
200
400
600
800
1000
1200
1400
1600
ever seen
special effects
waste time
looks like
dont know
ive seen
much better
one worst
look like
even though
ive ever
low budget
bad movie
im sure
watch movie
worst movie
see movie
dont think
pretty much
really bad

Trigrams Top 20 Trigrams IMDB DATA SET

Top 20 Trigrams (Trinity) in Positive Reviews:			Top 20 Trigrams (Trinity) in Negative Reviews:	
ive ever seen	348		ive ever seen	637
new york city	180		worst movie ever	352
world war ii	153		dont waste time	308
one best movies	139		one worst movies	301
one best films	123		movie ever seen	220
based true story	120		worst movies ever	200
first time saw	115		waste time money	170
first saw movie	91		one worst films	163
films ever made	89		dont get wrong	158
films ive seen	88		movie ive ever	158
seen long time	85		worst film ever	156
dont get wrong	82		movies ever seen	152
highly recommend movie	81		movie ever made	137
movie ever seen	79		worst movie ive	135
international film festival	79		bad acting bad	120
well worth watching	79		movies ive ever	116
would love see	79		complete waste time	112
first saw film	73		worst films ever	107
movie ive seen	71		doesnt make sense	103
many years ago	70		movie ive seen	102





Most common words or phrases in positive vs. negative reviews

IMDB data set :-

movie, film, one, like, even, good, bad, would, really, time, see, dont, get, much, story, people, could, make, made, movies, ever seen, special effects, waste time, looks like, dont know, ive seen, much better, one worst, look like, even though, ive ever, low budget, bad movie, im sure, watch movie, worst movie, see movie, dont think, pretty much, really bad, ive ever seen, worst movie ever, dont waste time, one worst movies, movie ever seen, worst movies ever, waste time money, one worst films, dont get wrong, movie ive ever, worst film ever, movies ever seen, movie ever made, worst movie ive, bad acting bad, movies ive ever, complete waste time, worst films ever, doesnt make sense, movie ive seen, great, well, also, first, films, love, best, one best, first time, new york, years ago, great movie, good movie, well done, real life, pretty good, must see, saw movie, new york city, world war ii, one best movies, one best films, based true story, first time saw, first saw movie, films ever made, films ive seen, seen long time, highly recommend movie, international film festival, well worth watching, would love see, first saw film, many years ago

Data set comparison

Feature	**IMDB**
Domain	Movie reviews
Size	50,000 labeled reviews
Language	English
Text Length	Long-form reviews (100–300+ words)
Label Type	positive, negative
Structure	2 columns: review, sentiment
Preprocessing Needed	Basic HTML removal, lowercasing

Aspects Identify from These Datasets

IMDB Reviews (Long-form):

- **Aspects**:
    - Acting, direction, storyline, cinematography, music, screenplay
    - Each aspect can carry different sentiment → suitable for **Aspect-Based Sentiment Analysis (ABSA)**

Comparison of the Two Datasets and Highlight Key Differences

Category	**IMDB Reviews**
Text Type	Long-form, structured reviews
Tone	Formal/neutral
Content	Movies and storytelling

Category	IMDB Reviews
Noise Level	Relatively clean
Preprocessing Complexity	Medium
Label Balance	Balanced (~25k each)
Use Cases	Product/movie reviews, recommendation systems
Aspect-level Analysis	Easier (explicit mentions of plot/acting etc.)
Generalization Ability	Suitable for fine-tuned sentiment tasks