Sarcasm Detection in IMDb Movie Reviews

Business Problem

Online reviews hold immense power in the film industry, shaping audience perception and box office success. However, sarcasm, a common weapon in the reviewer's arsenal, can wreak havoc on sentiment analysis systems. These automated tools struggle to decipher sarcastic intent, leading to misinterpretations that can skew audience ratings, mislead studios, and ultimately, disappoint moviegoers.

Consider these examples:

1. **Positive Review:** "Interesting, good filmmaking, surprisingly authentic." (Clear and positive)

2. **Negative Review:** "Dialogue is atrocious, acting phoned-in. Big disappointment." (Stronger wording emphasizes negativity)

3. **Sarcastic Review (Difficult to Detect):** "This movie was so bad, I'd rather gnaw on a golf ball." (Exaggerated statement hints at sarcasm)

4. **Sarcastic Review (Difficult to Detect):** "Wow, amazing plot! So many emotions! A true contender for Best Movie That Will Put You To Sleep!" (Sarcastic - Negative)

While the first review is clearly positive, the second one's negativity is easy to understand. However, the third and fourth review poses a challenge. The sentiment analysis system might interpret it literally, classifying it as negative in case of third review and positive in case of fourth review, when it's actually a sarcastic jab highlighting the movie's awfulness.

This is where our project comes in. We're developing a sarcasm detection system specifically designed for IMDb reviews. By identifying these witty critiques, our system will improve the accuracy of sentiment analysis, providing more reliable data for both audiences and the entertainment industry. With this enhanced understanding, moviegoers can make informed decisions, and studios can gain valuable insights into audience preferences.

Proposed Solution

Our proposed solution is to build a machine learning model capable of identifying sarcasm in IMDb reviews. The approach involves several key steps:

1. **Data Collection**: Gather a large dataset of Amazon reviews.

2. **Data Preprocessing**: Clean and preprocess the text data to ensure it is suitable for model training.

3. **Feature Engineering**: Extract relevant features that can help in detecting sarcasm, such as linguistic cues and contextual information.

4. **Model Training**: Train various machine learning models, including both classical algorithms (e.g., SVM, Naive Bayes) and advanced neural networks (e.g., LSTM, BERT).

5. **Model Evaluation**: Evaluate the models using appropriate metrics (e.g., accuracy, F1-score) to determine the best-performing model.

6. **Deployment**: Integrate the model into an existing sentiment analysis pipeline to improve overall sentiment detection accuracy.

Scope of the Solution

The scope of this project includes:

In-scope:

- Developing a sarcasm detection model specifically for English-language IMDb reviews.
- Evaluating model performance using a curated dataset of IMDb reviews.
- Focusing on text-based reviews, excluding multimedia content.

Out-of-scope:

- Detecting sarcasm in non-English reviews.
- Handling non-textual sarcasm (e.g., in images or videos or memes).
- Real-time detection of sarcasm during review submission.

Various Datasets Considered

We considered several datasets for our project:

1. **Amazon Customer Reviews (Amazon Product Review Dataset):**

Characteristics: A large dataset of Amazon product reviews, including review text, ratings, and metadata.
Pros: Rich and diverse in terms of product categories and review content.
Cons: Not specifically labeled for sarcasm, requiring manual or automated labeling.
Ref: https://github.com/ef2020/SarcasmAmazonReviewsCorpus

2. **SARC: A Corpus of Sarcastic Tweets:**

Characteristics: A dataset of tweets labeled for sarcasm.
Pros: Explicitly labeled for sarcasm.
Cons: Social media language and context differ significantly from product reviews, reducing applicability.
Ref: https://data.mendeley.com/datasets/fn2mmff85g/1

3. **The Sarcasm Corpus V2:**

Characteristics: A dataset containing sarcastic and non-sarcastic sentences from news articles.
Pros: Labeled for sarcasm.
Cons: Context and language style differ from those in Amazon reviews.
Ref: https://www.kaggle.com/datasets/coldn00lde s/sarcasm-corpus-v2oraby-et-al

4. **IMDb Movie reviews** :

Characteristics: The dataset consists of a rich archive of text reviews for movies, gathered from IMDb.
Pros: Labeled as positive or negative review
Cons: We're enriching our dataset by manually classifying reviews as sarcastic or non-sarcastic using available tools
Ref: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Final Dataset Selection

We selected the IMDb Reviews dataset for the following reasons:

- **Relevance:** The reviews are directly from IMDb, ensuring the context and language used are highly relevant to our target application.
- **Volume:** The large volume of reviews provides sufficient data for training robust machine learning models.

Use case of the sarcasm detection in movie reviews

- **Challenge:** Standard sentiment analysis models are good at identifying **clear** positive and negative reviews (e.g., "Great movie!" or "Awful acting!").
- **The Sarcasm Problem:** These models often struggle with sarcasm. A review like "Wow, best movie ever! (I fell asleep twice)" might be misinterpreted as positive, when it's actually negative sarcasm.
- **Our Solution:** We're building a **sarcasm detection model** specifically for movie reviews.
- **What it Does:** This model will analyze reviews that **confuse** sentiment analysis models. It will identify sarcasm and determine its true sentiment (positive or negative).
- **Working Together:** By combining a sentiment analysis model with our sarcasm detection model, we can create a **more complete** evaluation tool for online movie reviews.
- **Benefits:**

    - Moviegoers will have a **clearer understanding** of reviews, leading to informed decisions.
    - Studios will gain **valuable insights** into audience preferences.