IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Spoiler Detection Using Machine Learning

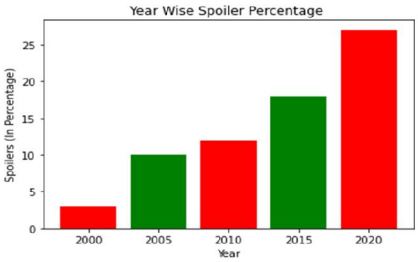
Atharv Kulkarni[1], Malhar Lohar[2], Manav Ahuja[3], Atharva Shastri[4], Prof. Yogesh Handge[5]
UG Student, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune [1,2,3,4]
Professor, Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune[5]

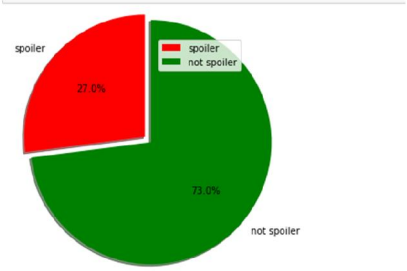
Abstract: *Over the course of the lockdown, television shows, web series, and movies have been viewed more than ever before. In order to pick a show or movie to start, we always need to browse through several reviews as it feels like a significant investment of our time, and we want to be sure that they will be genuinely entertaining and gauging. However, these reviews often contain information that reveal things about the plot that a viewer should not know prior to watch the respective show or movie. These bits of information are popularly known as spoilers and they possess the potential of greatly impacting a viewer's experience. These viewers may lose interest and in turn the production companies will to suffer from a loss of revenue. The issue at hand is that we want to read reviews before starting a movie or show, but we do not want to read a spoiler. How can we be sure that a review does not contain a spoiler? We need some warning, and that is why we are attempting to build a machine learning model that uses Natural Language Processing (NLP) to predict whether or not a particular review contains a spoiler, hence, serves as a warning. We will be using an IMDB review dataset to train and test our model.*

Keywords: Machine Learning, Neural Network, Deep Learning, LSTM, Natural Language Processing, etc.

I. INTRODUCTION

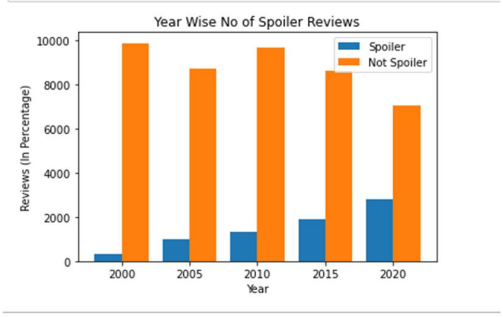
Most people would probably say that the Entertainment industry is a frivolous one. It holds no importance when it comes to real-world issues. Indeed, it isn't curing cancer or building bridges. However, Entertainment is vital in our lives. A Stan Lee quote describes this idea perfectly, "Entertainment is one of the most important things in people's lives. Without it, they might go off the deep end." Not only does entertainment keep humankind sane, but these industries around the world are also usually worth trillions of dollars. They have considerable contributions to their countries' economies indirectly. They trigger chains of activities like demand for tourism, accommodation facilities, food items, transportation services, health services, and local labor. Movies influence us and, in some cases, even consume us.





IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

II. DATA DESCRIPTION



The dataset is fetched from https://www.kaggle.com/rmisra/imdb-spoiler-dataset, it contains: movie details and user reviews.First dataset (containing movie details) has a total count of 1,572 movies, movie_id = primary key, duration = runtime of the movie.

	plot_summary	duration	genre	rating	release_date	plot_synopsis	title_x
tt0126886	Tracy Flick is	1h 43min	['Comedy', 'Drama', 'Romance']	7.3	07-05-1999	Jim McAllister (Matthew Bro	Election
tt0090605	57 years after Ellen	2h 17min	['Action', 'Adventure', 'Sci-Fi']	8.4	18-07-1986	After the opening credits, w	Aliens
tt0143145	James Bond is	2h 8min	['Action', 'Adventure', 'Thriller']	6.4	19-11-1999	In Bilbao, Spain, MI6 agent,	The World Is Not Enough
tt0082096	It is 1942 and the	2h 29min	['Adventure', 'Drama', 'Thriller']	8.4	10-02-1982	The movie opens with a scre	The Boat
tt0499448	A year has passed	2h 30min	['Action', 'Adventure', 'Family']	6.6	16-05-2008	On a cloudless night in Narn	The Chronicles of Narnia: Pr
tt0475784	Westworld isn't you	1h 2min	['Drama', 'Mystery', 'Sci-Fi']	8.9	29-11-2016	Ford brings Lowe back in the	Westworld
tt0120363	While Andy is away	1h 32min	['Animation', 'Adventure', 'Comec	7.9	24-11-1999	As the movie begins, Woody	Toy Story 2
tt0109707	Because of his	2h 7min	['Biography', 'Comedy', 'Drama']	7.9	07-10-1994	The film opens with a thund	Ed Wood
tt1217209	Set in Scotland in a	1h 33min	['Animation', 'Adventure', 'Comec	7.2	22-06-2012	In a prologue, we see Lord	Brave
tt0140352	Balls-out "60	2h 37min	['Biography', 'Drama', 'Thriller']	7.9	05-11-1999	In Lebanon, Hezbollah milita	The Insider
tt0351283	At New York's	1h 26min	['Animation', 'Adventure', 'Comec	6.9	27-05-2005	At New York City's Central P	Madagascar

Dataset 1: Movie Details

review_date	movie_id	user_id	is_spoiler	review_text	rating	review_summary
22-May-05	tt0080684	ur2521222	TRUE	SPOILERS Three years after "Star Wa	10	finest of the trilogy because Lucas didn't write the script
03-Apr-04	tt0080684	ur1173088	TRUE	(Note: This review contains some	10	The best "Star Wars" film ever made.
13-Aug-99	tt0080684	ur0415718	TRUE	While the Titanic made all the rich pe	10	Better than Titanic!!!
02-Feb-06	tt0080684	ur6643268	TRUE	"Star Wars: Episode V - The Empire S	10	The best movie of all Star wars
05-Aug-06	tt0080684	ur11569016	TRUE	I am A Star Wars fan, always have be	9	There's something in the way
21-Jan-14	tt0080684	ur50065462	TRUE	This movie is my all-time favorite. Fro	10	My favorite movie of all time.
26-Jan-08	tt0080684	ur15327507	TRUE	And I've seen some great films. Howe	10	The only film that gets a 10
29-May-05	tt0080684	ur4589082	TRUE	The best of the Star Wars series? Wit	10	If I could rate it above 10, I would.
27-Sep-15	tt0080684	ur53403649	TRUE	In many ways, The Empire Strikes Bac	9	The perfect sequel to its perfect predecessor
27-Mar-15	tt0080684	ur59488753	TRUE	Although Star Wars Episode IV was gr	10	The absolute best!

Dataset 2: Review Details

While the second dataset (containing details about user reviews) has a total 5,73,913 reviews posted by 2,63,407 users. Among all the records, 1,50,924 reviews contain spoiler content. Each review is labeled with true and false value depending upon the spoiler content in the review. Only the plot_synopsis having more than 50 words are selected for the training purpose.is_spoiler = spoiler (true) or not (false). review_text may contain the spoiler.Empty plot_synopsis entries. Some movie IDs have more than 4000 reviews while some of them have reviews less than 50. To reduce the variation in the count of the reviews we will filter out only those movie IDs which contain a review count between 300 to 500.

1. **Long short-term memory (LSTM)**

   Long short-term memory is an artificial recurrent neural network (RNN) architecture used in the field of Natural Language Processing and Deep Learning. Standard feedforward neural networks such as CNN and RNN, don't have feedback connections. LSTMs have an advantage in this aspect over simple neural networks. LSTMs can process single data points such as images as well as entire sequences of data such as speeches. The standard LSTM unit is made up of a cell, input gate, output gate and forget gate. The cell remembers the numbers

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

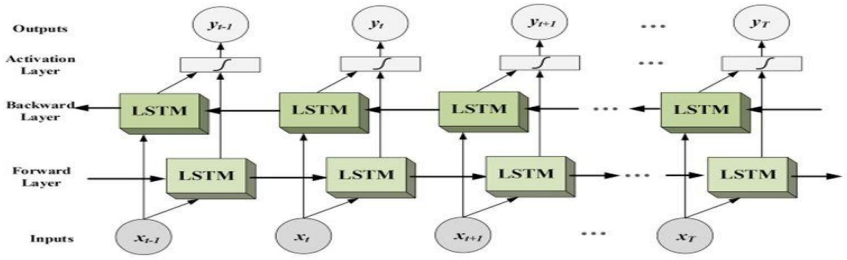
periodically and three gates control the entry and exit of information from the cell. LSTMs have been established to address the vanishing gradient problem that can be encountered when training traditional RNNs. They can remember the information for long periods of time.

2. **Bi-directional Long short-term Memory (Bi-LSTM)**

Bi-LSTM is the improved version of Long short-term memory (LSTM). This neural network has the capability to have the sequence information in both directions backwards (future to past) or forward (past to future). Using Bi-LSTMs we can stream the input to both directions to preserve the future and previous information.

In cases where all input time sequences are available, Bidirectional LSTMs train two instead of one LSTM input sequence. The first is in the correct order sequence and the second is a revised copy of the input sequence. This can provide more context in the network and result in faster and more efficient learning.

This type of architecture has many advantages in real-world problems, especially in NLP. The main reason is that every component of an input sequence has information from both the past and present. For this reason, Bi-LSTM can produce a more meaningful output, combining LSTM layers from both directions.



III. Literature Survey

(Sreejita Biswas and Goutam Chakraborty). The goal of their research is to identify spoilers in movie reviews to prevent ruining the movie watching experience of the readers. They extracted the imdb-spoiler-dataset (contains two parts, movie details and user reviews) from Kaggle and applied standard text cleaning steps like removing stop words, converting all words to lower-case and stemming. Their approach was to compare the synopsis and the movie review, the more similar they are the more likely it is that a spoiler is present in the review. Using the Random Forest Classifier with 'Gini' criterion they were able to achieve an accuracy of 86 percent with their model.

(Hwanjo Yu Sungho Jeon, Sungchul Kim). The goal of this research paper was to detect specifically those spoilers that appear on Social Networking Sites. They aimed to build a spoiler detection model which would allow people to freely go through social media and stay connected with their friends whilst still avoiding any spoilers. They generated their own dataset by collecting tweets regarding season 13 of Dancing with the Stars US and the 2014 World Cup. They found four features that can help in distinguishing tweets with spoilers from tweets without spoilers: Named Entity, frequently used verb, objectivity and URL and the tense of the tweet. To save time they chose a semi-supervised learning approach and trained their model using the SVM classifier algorithm. They were able to achieve an f-score of 0.7912.

(Saarthak Sangamnerkar Allen Bao, Marshall Ho). Using natural language processing to detect spoilers in book reviews.} – The aim of this research paper was to detect spoilers in book reviews. They used the UCSD Goodreads Spoiler dataset to train their model. They experimented with LSTM, BERT and RoBERTa language models. LSTM gave them the most promising result with an AUC score of about 0.91.

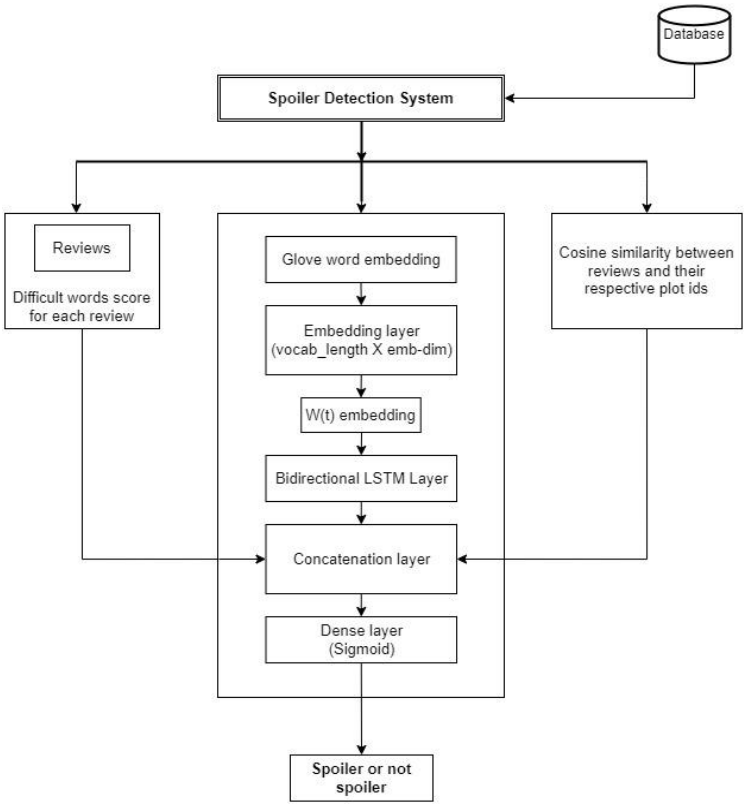
IV. Methodology

The proposed Spoiler detection model is based on Bi-directional LSTM neural network. Plot synopsis consists of entire movie details and can be considered as the ultimate spoiler. Movie reviews from the IMDB dataset are first pre-processed. Punctuation and stop words are removed from reviews and plot synopsis. Synopsis and movie reviews are turned into space-separated padded sequences of words. These sequences are again further split into lists of tokens using tokenizer.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Global Vectors for Word Representation (GloVe) embeddings is provided by the Stanford NLP team [5]. GloVe embedding are pre-trained and are used to deal with high-volume corpus. The embedding layer will load the weights from GloVe instead of loading random weights.

Once the text pre-processing is done, the approach is to score the similarity between the plot and the reviews. Features such as similarity score and difficult words score are extracted. Padded reviews are passed to Bi-directional neural network having embedding layers and dense layers. The model is trained repeatedly to reduce work loss and improve accuracy. Binary cross-entropy is used to detect reviews containing spoilers and non-spoilers. Adaptive learning optimization algorithms such as RMSProp and Adam are used to improve the model performance.



V. RESULTS

All codes are written in Python 3.8, using Tensorflow 2.0. All experiments have been performed on a Core™ processor Intel® CPU i7-4790U 3.60 GHz with 16 GB RAM. The two datasets utilized in this study are obtained from an open Machine Learning Repository accessible at Kaggle (refer Data Description section). The performance accuracy of the candidate model can be calculated using various available evaluation metrics. To check the prediction accuracy whether the news article is correctly classified as SPOILER or NON-SPOILER, accuracy as evaluation metric is used. Accuracy is calculated as ratio of number of correctly predicted samples to total number of samples for a given data set. The datasets are partitioned as 78:2:20 to train, validate and test the model.

Model	Accuracy
Bi-directional LSTM-RNN	77%

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

VI. Conclusion

In our implementation using the LSTM language model, we achieved an accuracy of 78 percent. While there is room for improvement, this model still detects the majority of spoilers, saving readers from having their movie-watching experience ruined and production companies from having their potential for generating revenue disturbed. Despite having the feature to tag spoilers, most reviewers do not use it. Hence, review platforms like IMDB and Rotten Tomatoes could use the proposed system to have reviews appropriately tagged automatically.

References

[1] Sreejita Biswas and Goutam Chakraborty. Movie reviews: To read or not to read! Spoiler detection with applied machine learning. In SAS GLOBAL FORUM. SAS, 2020.
[2] Hwanjo Yu Sungho Jeon, Sungchul Kim. Spoiler detection in tv program tweets. In Proceedings of the International AAAI Conference on Web and Social Media, pages 220–235. Information Sciences, 2016.
[3] Saarthak Sangamnerkar Allen Bao, Marshall Ho. Spoiler alert: Using natural language processing to detect spoilers in book reviews. https://arxiv.org/ pdf/2102.03882.pdf, 2021.
[4] IMDB Spoiler Dataset - https://www.kaggle.com/rmisra/imdb-spoiler-dataset
[5] GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glove/
[6] Hochreiter, J. Schmidhuber. (1997). “Long Short-Term Memory”, Neural Computation, 9(8):1735-1780.
[7] Has¸im Sak, Andrew Senior, and Franc¸oise Beaufays, “Long shortterm memory based recurrent neural network architectures for large vocabulary speech recognition,” arXiv preprint arXiv:1402.1128, 2014.
[8] Mike Schuster and Kuldip K Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997. [9] Alex Graves, Navd.
[9] Xiangang Li and Xihong Wu, “Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition,” in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 4520–4524.
[10] Felix A Gers, Nicol N Schraudolph, and Jurgen Schmidhuber, “Learning ¨ precise timing with LSTM recurrent networks,” The Journal of Machine Learning Research, vol. 3, pp. 115–143, 2003.