INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

www.ijraset.com

Call: ○08813907089 | E-mail ID: ijraset@gmail.com

Spoiler Detection & Early Spoiler Avoidance

Kunal Umaji[1], Vaibhav Mutange[2], Avanish Ramana[3], Safal Chayal[4], Shikha Sharma[5]

[1, 2, 3, 4]*Computer Department Trinity Academy Of Engineering,* Pune, India

[5]*Professor, Trinity Academy Of Engineering*Pune, India

Abstract: Spoiler detection is a major problem of the modern world as the internet is an imperfect entity which is still unable to block the spoiler, thereby killing the joy of a particular person who does not want to know about the particular information about movies, series, book but he gets to know about the important stuff about the things mentioned above before even watching it. So the spoilers that are present on the internet should be blocked. The various Machine Learning algorithm or Natural Language Processing can be used for detecting the spoilers as Machine Learning is the most advanced technology and previous researches have proved the usefulness and importance of theMachine Learning in various fields and in detecting the spoileras well.

Index Terms: Spoiler, Machine Learning

I. INTRODUCTION

Information that gives the major details about the entertain- ment can be called as spoilers. Detecting the spoilers hasalways been a difficult task as it is quite a confusing process because there are an uncountable number of texts availableon the internet and it is a lengthy task to detect those texts which are spoiler. The previous researches have proved that Machine Learning algorithms are very efficient in terms oftext processing, and as spoilers are nothing but the texts that contain some information that the user should not have seen bythe user. So with the help of Machine Learning, we generatean algorithm which automatically detects a spoiler and block them. But it is very important to define the terms or quantity on which we are going to detect the spoiler. So what we need is an automated technique to determine the spoiler and hence, as it will automatically detect the spoiler, this task can be doneusing Machine learning. To go in more depth about the spoiler detection, it is very important to understand the concept of text classification because spoiler is nothing but just a text and there are a lot of works has been done on text classification. Text classification is a machine learning technique that automatically assign a tagor category to the text, so in this case of spoiler we will assigna text as a spoiler if it fails to pass some metrics that has been defined, so it will be considered as a spoiler. There are various algorithms that can be used in this work that are we goingto discuss in this research paper. So it will give us the idea about accuracy if every algorithms and we will get to choose the most suitable algorithm that gives the most accuracy and more reliable than others and will give the clear idea about detecting and blocking the spoilers which are available on the internet.

II. LITERATURE SURVEY

Graph Neural Network can be used effectively to find the spoilers and have the accuracy of 87 percent which is highest among all the previous methods, and this model is calledas SDGNN and has attention mechanism that helps it to achieve this much of perfection [7]. Graph Neural Networkis a neural network that can be applied to graph theory andthe neural network is a set of algorithms that work exactlylike the human brain. Thus, Graph Neural Network worksefficiently in detecting spoilers as it uses the Neural network on the graph.

The whole spoiler blocker is not feasible to apply for overallinternet as for some people it will not be considered as spoiler but for some it is, so user-specific web filtering can be aone good option in which the user will be able to filter content based on his preference [6]. So the user can definethe definition for the spoiler of his own that can be movie name, book name, web series name, and by adding this intothe model the corresponding texts which are doubtful will be blocked from client machine. This method has been used in spoiler protection 2.0, which is an extension for chrome in which user adds keywords and the malicious links, texts that can be spoiler and any doubtful resource gets blocked and thisis how the user can surf the internet without any worries of seeing any spoiler.

In 2016, the basic model for the spoiler detection has been implemented that has been tested on the twitter platform whichblocks the spoiler. Twitter is one of the most famous platformsthat has been blamed for generating spoilers as there are continuous tweets are posted and that tweet can also contain spoilers. So Sungho Jeon, Sungchul Kim, and Hwanjo Y has implemented a model for the twitter using Support Vector Machine algorithm that successfully block the spoiler. [5]

Avoiding spoiler while using internet is the common problem while using social media, so Jordan Boyd-Graber, Kimberly Glasgow, and Jackie Sauter Zajac has implemented supervisedclassifiers for detecting spoiler using the approach that nothingis spoiler as not everything available on the internet is not a spoiler assuming this made their model more reliable, their goal was simple that to take an example of spoiler and make the system that can determine a new sentence if it is spoileror not [1].

III. NOTABLE ALGORITHMS FOR DETECTING SPOILERS CLASSIFIERS

Classifiers are the algorithms that are very useful in terms of classifying something in this work we are going to classify whether the review is a spoiler or not. Sowe are going to see the best classifiers one by one.

1) *Naive Bayes Classifier:* Naive Bayes classifiers use probability to Predict whether an input will fit into a certain category. The Naive Bayes algorithm family includes a range of different classifiers Based on a theorem of probability.These classifiers can determine The probability of an input fitting into one or more categories. In multiple category sce- narios, the algorithm reviews the probabil- ity that a data point fits into each classification. After comparing The probabilityof a match in each category, it outputs the category That is most likely to match the given text.

2) *Decision Tree:* A decision tree is a classificationalgorithm that uses a process of division to split data into increasingly specific Categories. It's called a decision tree because the classification Process resembles a tree's branches when represented graphically [3]. The algorithm works on a supervised model and requires high- Quality data to produce good results. Since the primary goal of a decision tree is to make increasingly Specific distinctions, it has to continuously learn new classification rules. It learns these rules by applyingif-then logic to training data. The algorithm continues theclassification process until it reaches a designated stopping condition.

Bayes' Theorem

Conditional probability: Probability of a label y occurring as a result of the given features x.

Joint probability: Probability of features x occuring together with label y.

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

Marginal probability: Probability of features x occurring across all possible labels y.

Fig. 1. Formula
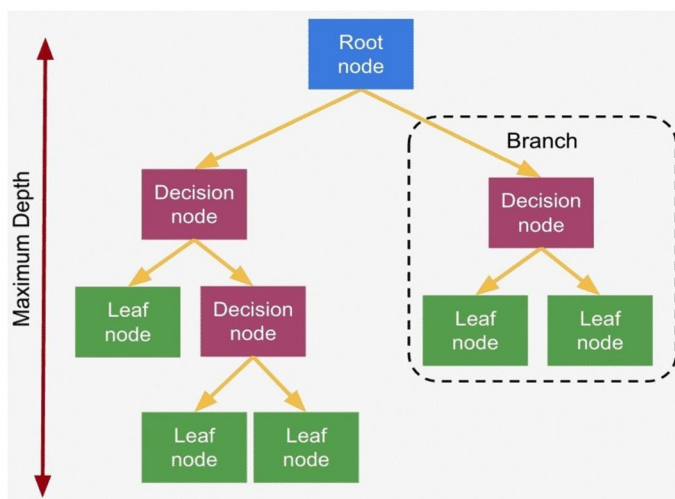

Fig. 2. Decision Tree

3) *Support Vector Machine:* A support vector machine (SVM) is a simple algorithm that professionals can use for classification or regression activities. They work by finding a hyperplane within a data distribution, which you can visualize as a line separating two different classes of data. There are often many hyperplanes capable of separating the data, andthe algorithm will select the optimum line of separation. Inthe SVM model, the optimum hyperplane is the dividing line that offers the greatest margin between the different classes. SVM are capable of working in more than one dimension if they are unable to find an ideal hyperplane to separate the data into two dimensions. This makes them extremely effective for creating classifications from complicated data distributions.
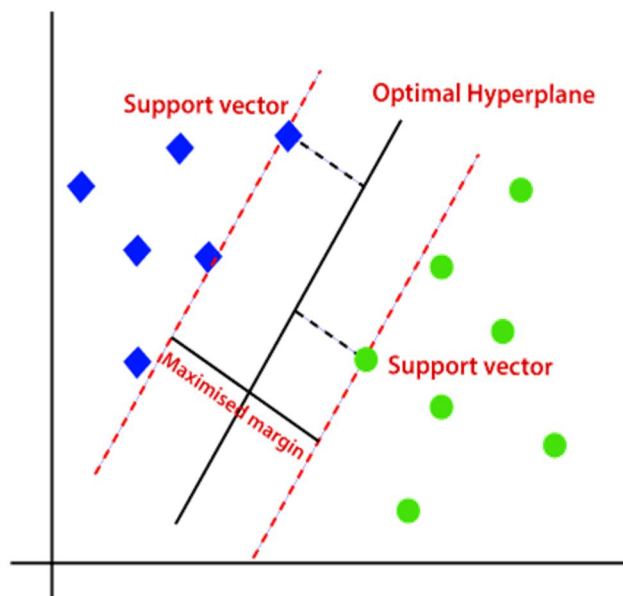

Fig. 3. SVM

The larger the complexity of the data inputs are, the more accurate the SVM becomes, thereby making them excellent machine learning tools.

IV. METHODOLOGY

To detect whether the review contains a spoiler or not, it is very important avail the data, for the reviews of the movie,we can take the data-set from the kaggle which contains 573,913 reviews and other details such as movie id, plot summary, duration, genre, rating, and plot synopsis. Once the data cleaning and the pre-processing is done we can attemptto predict whether the review is a spoiler or not on the basis of plot synopsis and review text that is given in the data-set. By applying various text processing techniques we can findthe correlation between the reviews and plot synopsis, it can give us an idea whether the review is spoiler or not [2].

This is how the whole process is supposed to look like once themodel has been made, trained, and it's tested. The model will take a review as input and will predict accordingly whetherthe review is a spoiler or not. If the review is not a spoilerthen our model will allow to post that review; otherwise the review will get hidden or we can add warning for the reviewas "Spoiler Here". Model has to be trained on the large data- set as the review can be of different types and some can be spoiler while not all of them, so we must be careful because our model can give the false spoiler as well. hence it has be trained on the large data-set, testing also plays very important part. Once the training, and testing has been done properly, then only it will work properly. Also, the data-set that we are going to use has to go through data-pre-processing techniques as to remove the unnecessary data from the data-set, the noise,for aligning the data in the proper format. Once this is done now it is very important to select the features as our predictionswill be based upon the features. Features are those aspects on which we are going to predict the reviews, if they are spoilers or not.
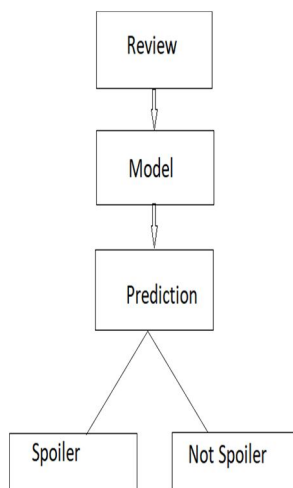
Fig. 4. Process

Also for the text classification the BERT (BidirectionalEncoder Representations from Transformers) model is very powerful to use and can used in the spoiler detection as it helps the computer to understand the meaning of the ambiguous language with the help surrounding text to establish somemeaningful context. BERT model can be used for detecting spoiler [4] as it is already pre-trained with billions of words and a lot of book corpus, making it very important model to be use for spoiler detection. Bert takes a sequence of wordsas input and then the Self-attention layer is applied to every layer and the result is passed through a feed-forward network, and then to the next encoder, this is what the model work flowlooks like.

Below is the visual representation of the detecting spoiler. The model is already pre-trained on a corpus of data to understand a language. Then it is fine-tuned with the desired data set that we are going to use so as to understand the data more clearly. Then it finally predicts what we want to know that is nothing but the spoiler. This is very convenient to use as pre-trained models are shortcuts which allows us to use knowledgeable models.
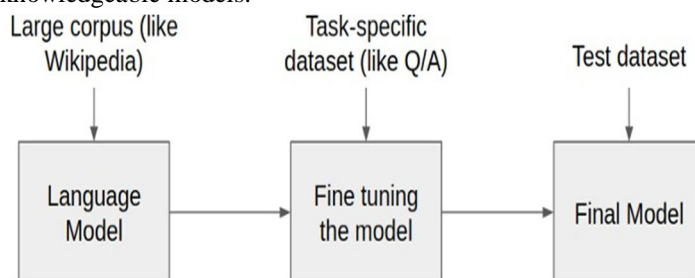

Fig. 5. Overview Of Bert

V. CONCLUSION

The advancement of spoiler detection is done by Machine Learning, Natural Language processing and it is still in progress as there is a lot of ongoing work and also everyone is trying to increase the accuracy as there is a long way to go in terms of increasing accuracy. So from This paperwe can conclude that Machine Learning Algorithms, and Natural Language Processing are being applied in almost everyplatform which tries to detect the spoiler and block themas they hold the many applications in text classification and spoiler is nothing but just a text and various classification techniques can be used to block the spoiler. Spoiler detection and Machine Learning have been bedfellows for many yearsas previous research shows. It has been a staple of detecting spoiler since day one and will never go away. As long asthere is Internet the spoilers will be there and there is no stopping them, it will be around, and as it will be around itis very important to stop the spoilers. Still, with the current technologies we are not able to get the accurate result as no model at the present has the accuracy more than 87 percent, and thus there is need of lot of work in the field of blocking spoilers because 87 percent is just not enough.

REFERENCES

[1] Jordan Boyd-Graber, Kimberly Glasgow, and Jackie Sauter Zajac. Spoiler alert: Machine learning approaches to detect social media posts with rev- elatory information. Proceedings of the American Society for InformationScience and Technology, 50(1):1–9, 2013.

[2] Buru Chang, Hyunjae Kim, Raehyun Kim, Deahan Kim, and Jaewoo Kang. A deep neural spoiler detection model using a genre-aware attention mechanism. In Pacific-Asia Conference on Knowledge Discoveryand Data Mining, pages 183–195.

[3] Springer, 2018. Sheng Guo and Naren Ramakrishnan. Finding the storyteller: automatic spoiler tagging using linguistic cues. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 412–420, 2010.

[4] Hans Ole Hatzel. Using neural language models to detect spoilers.

[5] Sungho Jeon, Sungchul Kim, and Hwanjo Yu. Spoiler detection in tv program tweets. Information Sciences, 329:220–235, 2016.

[6] Kyosuke Maeda, Yoshinori Hijikata, and Satoshi Nakamura. A basic study on spoiler detection from review comments using story documents. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pages 572–577, 2016.

[7] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. Fine-grained spoiler detection from large-scale review corpora. 2019.

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)