A Comparative Study of Machine Learning Approaches for Enhancing Copyright Protection Strategies

Audrey E. Ng

Abstract

In recent times, businesses have become more subject to copyright infringement due to the widespread accessibility of digital content and the ease of unauthorized reproduction and distribution. As a result, businesses lose customers and money to their plagiarized counterparts. In order to mitigate these effects, machine learning techniques can be utilized to detect plagiarized content. This paper aims to present and evaluate the effectiveness of various machine learning approaches in enhancing businesses' copyright protection strategies, including Support Vector Machines, Random Forests, and Convolutional Neural Networks.

Keywords: artificial intelligence, machine learning, copyright infringement, copyright protection, support vector machines, random forests, convolutional neural networks, intellectual property, plagiarism detection

Literature Review

Copyright protection is a crucial concern for businesses that create and distribute intellectual property. With the exponential growth of digital content and the rise of online piracy, traditional copyright protection methods are becoming insufficient. For instance, stamp-set designs can easily be copied and rebranded as a separate (and usually cheaper) product. Customers looking to buy the original stamp-set designs may be deceived by the copied products' design and attracted by the lower cost which may lead them to opt for the cheaper price instead. The traditional method of addressing these copied products is to manually analyze the similarities of the products and individually report those that do. However, with the rise of artificial intelligence, specifically machine learning, businesses now have the opportunity to address copyright infringement concerns in a more efficient manner. The following literature review attempts to understand the existing research and advancements in the field of copyright protection and machine learning.

Several studies have demonstrated the effectiveness of machine learning algorithms in detecting and identifying instances of copyright infringement. A study by Lelisa Adeba Jilcha and Jin Kwak (2022) emphasized the use of machine learning algorithms to detect advertisements served on suspected piracy websites. Just like original designs, online advertisements can easily be copied and reuploaded on a separate website that claims it to be their own. This action directly affects businesses and can cause billions of dollars lost in revenue. The proposed technique to address this issue involved the implementation of two models: support vector machines (SVMs) and a word2vec vectorization model. A support vector machine is a machine-learning algorithm that analyzes data to classify them into groups. The word2vec vectorization model takes words from a text and converts them into numerical vectors which represent the meaning or context of the words in a way that a computer can understand. The evaluation of this proposed technique found that their trained model can achieve an impressive 97% accuracy in identifying advertisement banners from piracy websites.

Similarly, a study by Kim et al. (2021) explores the use of a photo identification framework in protecting intellectual property from manipulations. The photo identification framework consists of various image-processing techniques that facilitate the extraction of distinctive features and patterns. The authors leveraged a training dataset with a wide range of authentic and manipulated images, which allowed their model to gain the ability to identify new differences between the two categories. The study also highlights the framework's adaptability across various industries and domains. Its potential applications extend beyond conventional product counterfeiting concerns to encompass image tampering in fields such as journalism and forensics; the versatility of their approach demonstrates that it is a useful tool in the ongoing battle against digital fraud and misinformation.

These studies demonstrate the potential of machine learning techniques in identifying instances of copyright infringement. From identifying copied designs to shielding brand reputation against counterfeit products, the applications of machine learning are already showcasing their capability to detect and counteract copyright violations. As businesses navigate the evolving landscape of intellectual property rights, utilizing machine learning techniques serves a dual purpose: preventing copyright infringements and cultivating consumer trust. This not only upholds fairness but also contributes to sustained competitiveness in the market. Additionally, the intersection of technology and copyright protection offers a key avenue for businesses to uphold their creative endeavors while preserving fairness and authenticity in their intellectual property.

Methodology

The traditional method of combatting copyright infringement is to manually identify plagiarized content and report each product individually, which can be tedious. To make this process easier, this research aims to present various machine learning techniques that could reduce the manual effort of identifying such content when implemented on a trained model. It is important to note that the solutions outlined in this paper are hypothetical and have not undergone empirical testing.

Support Vector Machines

By utilizing a machine learning algorithm called support vector machines (SVMs), a trained model can learn patterns and features from a labeled dataset consisting of both genuine and plagiarized content. The model can then establish a decision boundary that distinguishes specific categories.

The SVM's mechanism involves transforming data points into a higher-dimensional space, which allows the model to identify optimal hyperplanes that differentiate categories. This process is guided by "support vectors," key data points that are close to the decision boundary. The SVM then learns to generalize patterns and classify new data points.

The classifier function is defined as $f(x) = w \cdot x + b$. In this equation, $f(x)$ represents the classifier's prediction for the given data point x, w represents the weight vector associated with the decision boundary, and b represents the bias term, which is an adjustment that helps position the decision boundary in a way that improves how well it separates different categories. It is important to note that there is a lot more information regarding how a support vector classifier works. However, **Figure 1** depicts a general representation of the classifier function.

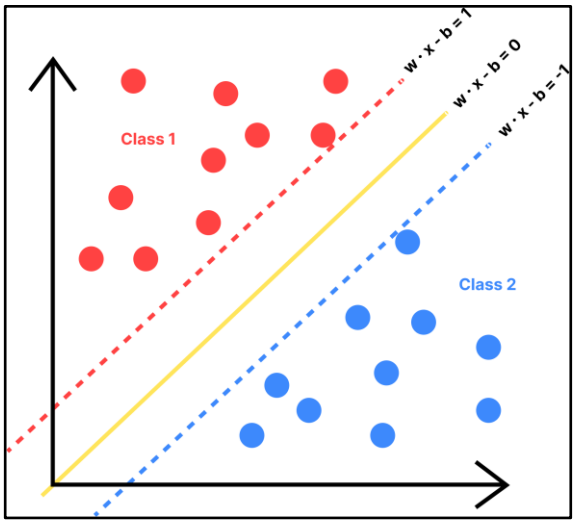

Figure 1. Graph of the support vector classifier function.

Random Forests

A random forest is a machine learning technique that involves constructing a group of decision trees that collaborate to make decisions. Each decision tree is trained on different subsets of a dataset. Random forests excel in their ability to handle diverse and complex datasets due to the use of multiple decision trees. Having multiple decision trees enhances generalization and mitigates the risk of overfitting, which occurs when a model performs poorly when given new data despite its strong performance on the training dataset.

As shown in **Figure 2**, The key process within random forests is the creation of individual decision trees and the subsequent fusion of their decisions. Each decision tree contributes its own insight by learning distinctive patterns from a subset of data and a random selection of features. The group of decision trees as a whole then integrates these insights to deliver a collective, final decision.

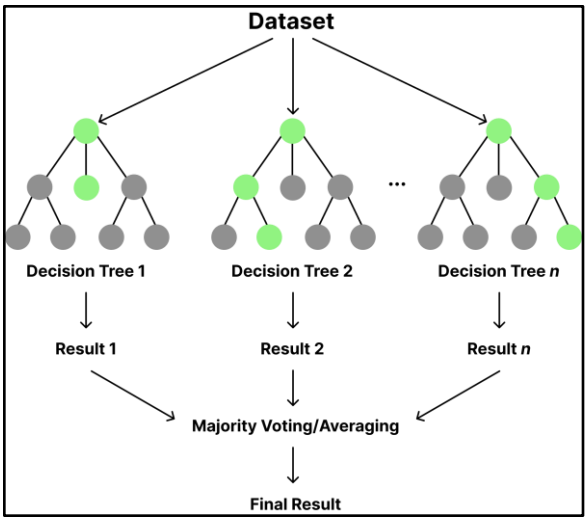

Figure 2. Random forest algorithm process.

Convolutional Neural Networks

A neural network is a machine learning model inspired by the connections between neurons in the human brain. Its framework comprises layers of interconnected nodes, or "neurons." These nodes process information and learn patterns from data.

A Convolutional Neural Network (CNN) is a type of neural network comparable to the human brain's ability to recognize patterns and features in visual stimuli since they start by using layers called "convolutional layers" which learn and extract features from visual data by applying small filters that detect patterns like edges and textures across it. Convolutional layers are followed by "pooling layers" that downsample the features, which reduce their spatial dimensions and in turn make the network more efficient. Finally, fully connected layers combine the learned features and make decisions. The hierarchical structure of CNNs allows them to learn simple features in early layers and more complex features in deeper layers, making them useful for tasks such as image recognition.

Discussion

This paper proposed several machine learning techniques that reduce the manual effort of identifying plagiarized content when implemented on a model. In this discussion, these techniques will explained regarding their benefits to aid businesses in the copyright protection process along with their potential limitations. Again, the solutions outlined in this paper are hypothetical and have not undergone empirical testing.

Support Vector Machines

As illustrated in **Figure 1**, a support vector classifier function has the ability to establish a decision boundary that differentiates between different classes. Therefore, in the context of copyright protection, SVMs can be trained using labeled datasets that contain both copyrighted material and potential infringements. Once trained, these models can classify new and unseen content while also flagging items that may require further human review.

Businesses can utilize SVMs to build automated systems that monitor digital platforms for unauthorized use of copyrighted material. For example, SVM-based algorithms can be employed to scan websites, social media platforms, and content-sharing platforms for instances of content that closely resembles copyrighted works. By setting appropriate thresholds, SVMs can help identify content that may be infringing on copyrights. Integrating SVMs would also save businesses the time and effort that would typically go into manually identifying infringing content.

However, it is important to consider the potential limitations of SVMs. If the dataset that an SVM is trained on is too large, the SVM could become computationally intensive and memory-demanding, which could delay the infringement detection process. If the dataset is unbalanced (too much of one type of content) or contains data that is incorrectly labeled, the SVM is likely to incorrectly classify new content. Therefore, if a business does decide to utilize an SVM, the dataset they use to train the SVM should be of a manageable size, balanced, and accurate.

Random Forests

Random forests can be trained using labeled datasets containing copyrighted and non-copyrighted content. Multiple decision trees allow random forests to capture complex patterns within the data, making them effective at distinguishing between different types of content.

Businesses can leverage random forests to build predictive models that determine whether a given piece of content is original or plagiarized. These models can consider a variety of features associated with the content, such as text, images, metadata, and more. The random forest algorithm is also capable of automatically selecting relevant features and learning complex relationships between them, which is particularly valuable when dealing with diverse forms of content that may have varying infringement indicators. Unlike SVMs, random forests are well-suited for handling unbalanced datasets, which are common in copyright protection scenarios since there may be more cases of infringing content than original content. They can also assign appropriate weights to different classes which ensure that both infringing and genuine content are accurately identified. This is crucial for reducing false positives and negatives in copyright detection. Since the random forest algorithm also includes a feature where a collective decision is made between the different decision trees errors determined by individual decision trees do not have much influence on the final decision. Thus, the risk of overfitting (when a model performs poorly when given new data despite its strong performance on the training dataset) is reduced.

While random forests can effectively handle datasets that are unbalanced and contain inaccurate classifications, one drawback is their computational complexity and memory requirements, especially with large datasets. Utilizing the random forest algorithm involves training multiple decision trees, which can become time-consuming and resource-intensive when processing large datasets. This could lead to slower response times in real-time monitoring scenarios and affect how fast businesses can identify and address violations.

Convolutional Neural Networks

Convolutional neural networks excel at image and pattern recognition, which is particularly useful for identifying copyrighted content and potential infringements within images and videos. By training CNNs on labeled datasets containing copyrighted material and non-copyrighted content, businesses can create models that learn distinct visual features, such as unique textures, shapes, and arrangements. Learning these features allows CNNs to identify subtle differences between original content and unauthorized reproductions, even when transformations or alterations have been applied. As with the other machine learning techniques, CNNs can enhance the efficiency of copyright monitoring by automating the process of content analysis which, in turn, saves businesses time and energy.

While CNNs offer remarkable capabilities for copyright protection, one limitation is the need for large amounts of labeled training data. Developing accurate CNN models requires extensive datasets containing both original and infringing content, which can be time-consuming and resource-intensive. This can pose challenges for businesses, especially those with limited access to diverse and well-labeled training content. Additionally, the complex architecture of CNNs can make them computationally demanding and require powerful hardware resources for efficient real-time analysis. Not every business has access to these resources.

Conclusion

This paper explored the potential of three machine learning approaches: Support Vector Machines (SVMs), Random Forests, and Convolutional Neural Networks (CNNs). These techniques offer different advantages in detecting and mitigating copyright infringement across various forms of content. SVMs excel in creating decision boundaries to classify different categories, while Random Forests leverage ensemble learning to handle complex and diverse datasets. CNNs have the ability to identify intricate visual patterns and features that distinguish original content from infringing content. While each of these machine learning techniques has its benefits, their implementation

comes with certain considerations: selecting appropriately sized and balanced datasets, dealing with computational demands, and ensuring access to powerful hardware resources for efficient real-time analysis. By adopting these techniques, businesses can enhance infringement detection, preserve brand reputation, and improve their consumer trust. These advancements offer a path towards more efficient and effective intellectual property safeguarding, especially in an era with rapid digital creativity and widespread content sharing.

References

Jilcha, L. A., & Kwak, J. (2022). Machine Learning-Based Advertisement Banner Identification Technique for Effective Piracy Website Detection Process. Computers, Materials, & Continua, 71(2), 2883-2899. https://doi.org/10.32604/cmc.2022.023167

Kim, D., Heo, S., Kang, J., Kang, H., & Lee, S. (2021). A Photo Identification Framework to Prevent Copyright Infringement with Manipulations. Applied Sciences, 11(19), 9194. https://doi.org/10.3390/app11199194