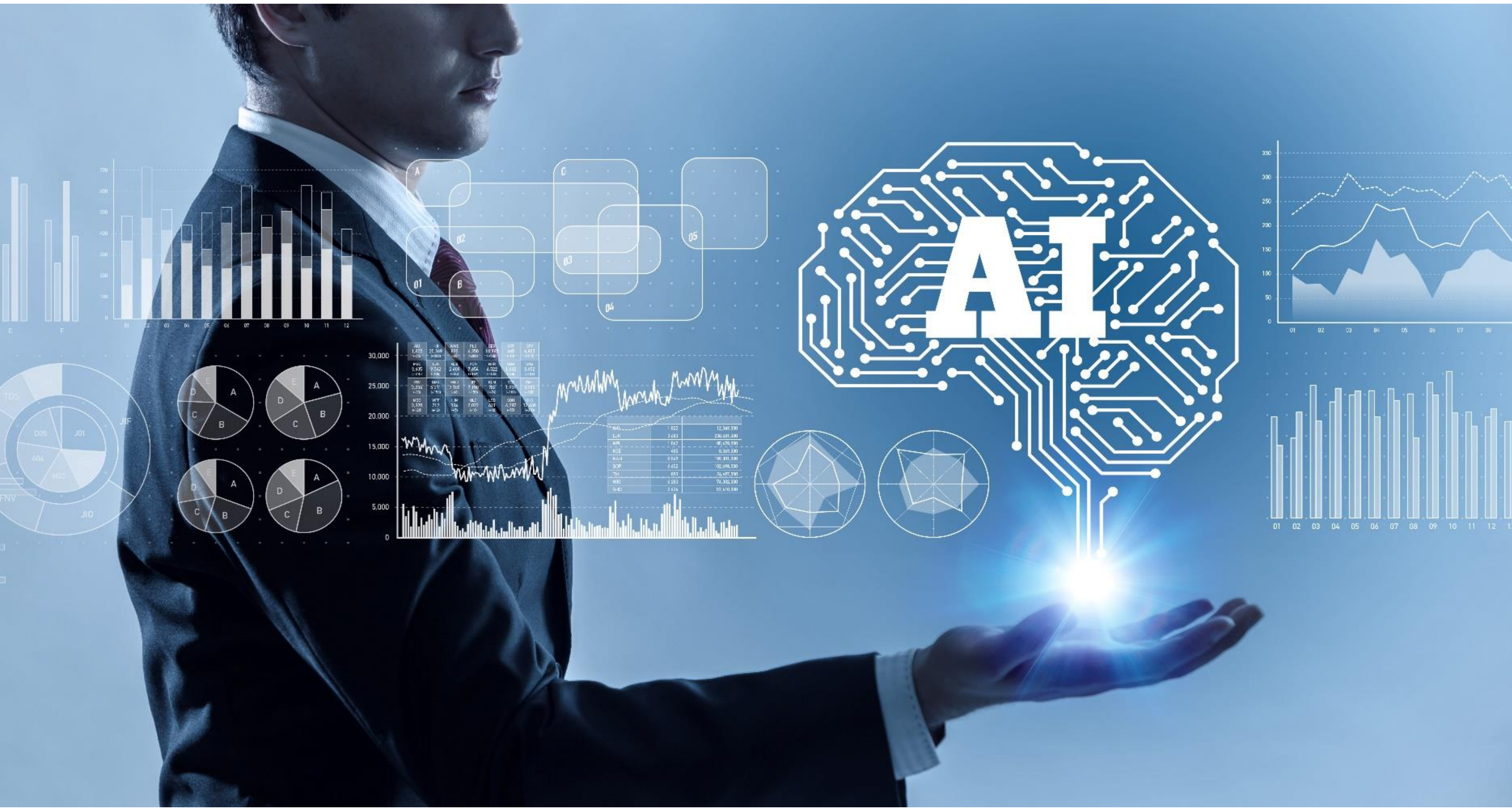
Upping the Game:

Using AI Natural Language Processing (NLP) To Proactively Detect Web Defacements



cloudsine

WEBORION

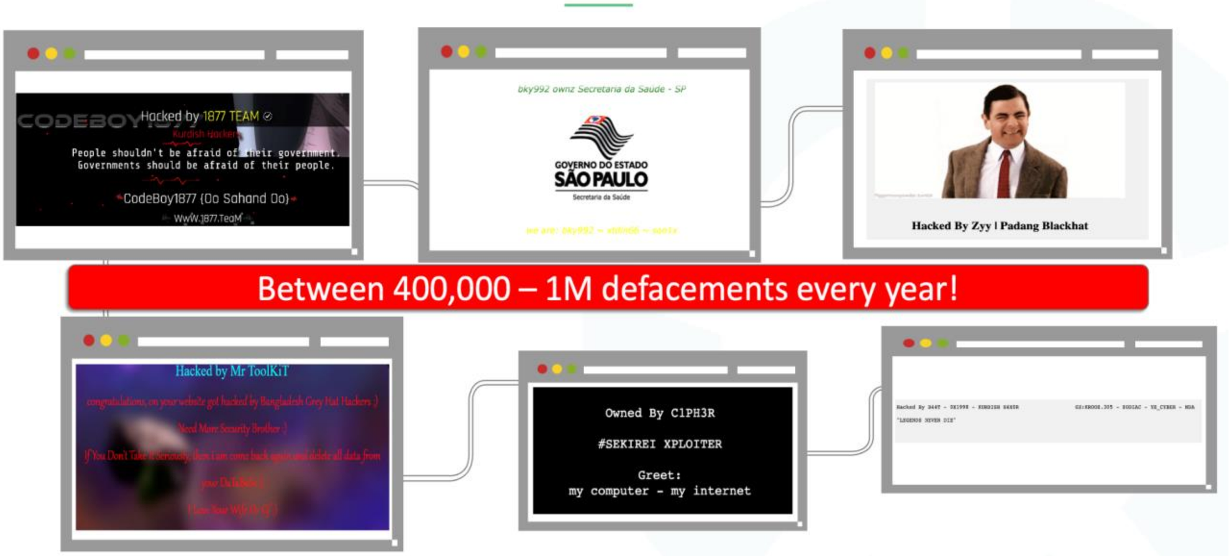
Introduction

With the rapid development of the internet around the world, the threat of website defacements is also increasing. According to Zone-h.org, about 500,000 websites were defaced in 2020, and the growth of e-commerce has only made the impact of website defacements even more damaging to an organization's reputation. The resulting loss in customer confidence often translates to declining revenues and hefty penalties by regulatory agencies if there is a data breach.

Several techniques can be employed to monitor the website defacements. However, most of these methods cannot distinguish between actual defacements by hackers and typical web page updates by administrators.

Leveraging on the power of Natural Language Processing (NLP), Cloudsine introduces a new defacement monitoring enhancement, the AI NLP Engine and integrated it with the current WebOrion Defacement Monitor. The new engine contains two different AI models, which have been named the Model T and Model X, respectively. Both models were trained with over 10 years of defaced datasets and can analyze the website HTML Text extracted, producing defacement confidence score that more accurately determines whether the webpage has been defaced with fewer false positives than ever before.

Why AI NLP?

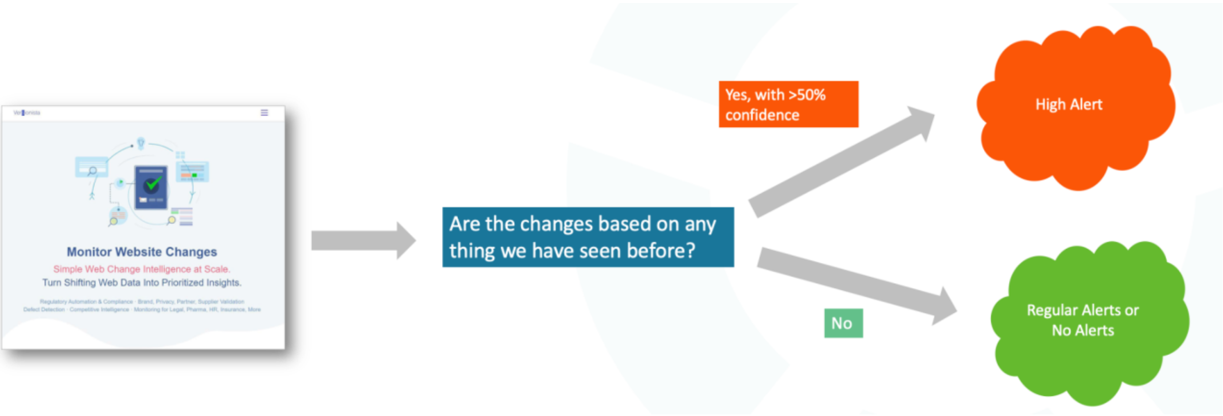

Hacked by 1877 TEAM
Kurdish Hackers
People shouldn't be afraid of their government.
Governments should be afraid of their people.
CodeBoy1877 (Oo Sahand Oo)
WwW.1877.TeaM
bky992 ownz Secretaria da Saude - SP
GOVERNO DO ESTADO SÃO PAULO
Secretaria da Saúde
Hacked By Zyy I Padang Blackhat
Between 400,000 – 1M defacements every year!
Hacked by Mr ToolKiT
congratulations, on your website got hacked by Bangladesh Grey Hat Hackers :)
Need More Security Brother :)
If You Don't Take it Seriously, then i am come back again and delete all data from your DataBase :)
Owned By C1PH3R
#SEKIREI XPLOITER
Greet:
my computer - my internet

Websites continue to be attacked because they are readily accessible on the internet for probing and hacking. Web defacements continue to be very prevalent, ranging from 400,000 to 1 million public websites being hacked globally every year. These range from government , enterprise , non-profit , small business websites, etc. There are significant business and security implications of defaced websites including reputational impact, loss of business and consumer trust, and defacements are often symbolic of deeper security issues of the website owners.

Historical observations of past defacements include the following:

- Defaced web pages can show hacker images, texts or even audio/video clips.
- Defaced web pages typically have black backgrounds, although quite few contain white or no background images at all.
- Some common keywords used in defaced web pages include “hacked,” “owned,” “pwned,” etc.
- Defacement text may or may not be in English.

We decided to use AI to learn from over ten million samples of defaced web pages over the past decade. This allows the WebOrion Monitor to add a new level of intelligence beyond just webpage change or keyword detection. We have also explored using both AI computer vision (CV) and text analysis techniques. Cloudsine’s researchers realized that as defaced webpages do not have representative color patterns, the use of text analysis is a better choice than CV. Thus, AI NLP was adopted to learn from past defacements and analyze and classify if the legitimacy of web page edits.


Monitor Website Changes
Simple Web Change Intelligence at Scale.
Turn Shifting Web Data Into Prioritized Insights.
Regulatory Automation & Compliance · Brand, Privacy, Partner, Supplier Validation
Defect Detection · Competitive Intelligence · Monitoring for Legal, Pharma, HR, Insurance, More
Are the changes based on any thing we have seen before?
Yes, with >50% confidence
High Alert
No
Regular Alerts or No Alerts

About Natural Language Processing (NLP) and Considerations Used in Our Models

Natural Language Processing (NLP) is a branch of Artificial Intelligence or AI. The technology enables computers to process and understand human language in the form of text to understand its full meaning. The field of statistical NLP combines computer algorithms with machine learning and deep learning models to automatically extract and classify elements of text and then assign a statistical likelihood to each meaning of those elements. Here NLP was utilized to process the text content of the websites, extract the embedded information, and make predictions to classify whether the websites are defaced or not.

The AI NLP engine has been built with the following considerations during its research and development:

- **Language Detection:** Language probability is analyzed, and the proportion of English is verified in a sample HTML text content. If there is a major change in language in the monitored webpage (e.g. from English to non-English), we will trigger alerts as webmasters typically do not modify the language of a webpage.
- **Stop words**: Most Stop words are removed from processing to enhance the accuracy of the AI model.
- **Syntax and Semantics:** Beyond just detecting keywords, the AI model will learn from the syntax and semantics of the sentences.
- **Transformer Attention Model (TAM) based training for NLP**: WebOrion's Model-T leverages on a proven AI NLP model with additional testing and enhancements with defaced data to provide the optimal accuracy to classify defacements.

About WebOrion's AI NLP Engine

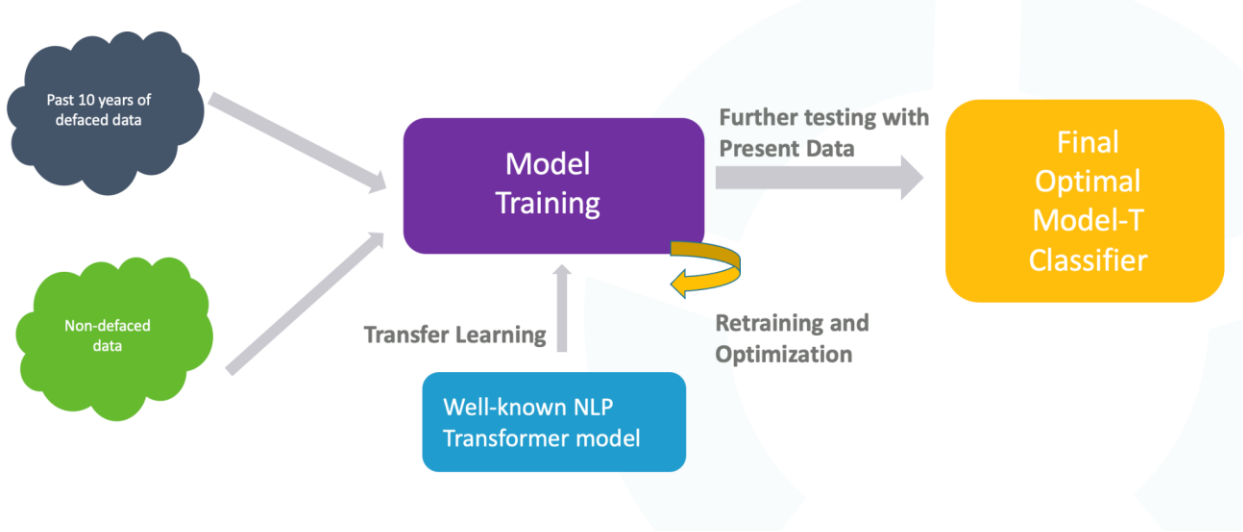
The idea here is to integrate the new NLP Engine with the current WebOrion platform, which is already widely used to check thousands of webpages for potential defacement attacks. The NLP technique helps analyze the text content on the target website instead of monitoring the simple changes found in websites. When the user adds a new website to WebOrion Defacement Monitor, an intelligent baseline is run three times on the target website. If any changes are detected, the integrity engine forwards the website to the AI NLP Engine for a further analysis. The model in the NLP Engine will then analyze the content of the websites independently and give the analysis result in the form of confidence between 0% and 100% to indicate whether the website has been defaced is defaced or not. A lower confidence score indicates that the model believes that the detected change to the web page is less likely to be a defacement. Conversely, a higher the confidence score indicates a high probability that a malicious webpage defacement has occurred.

Model T Training and Inference

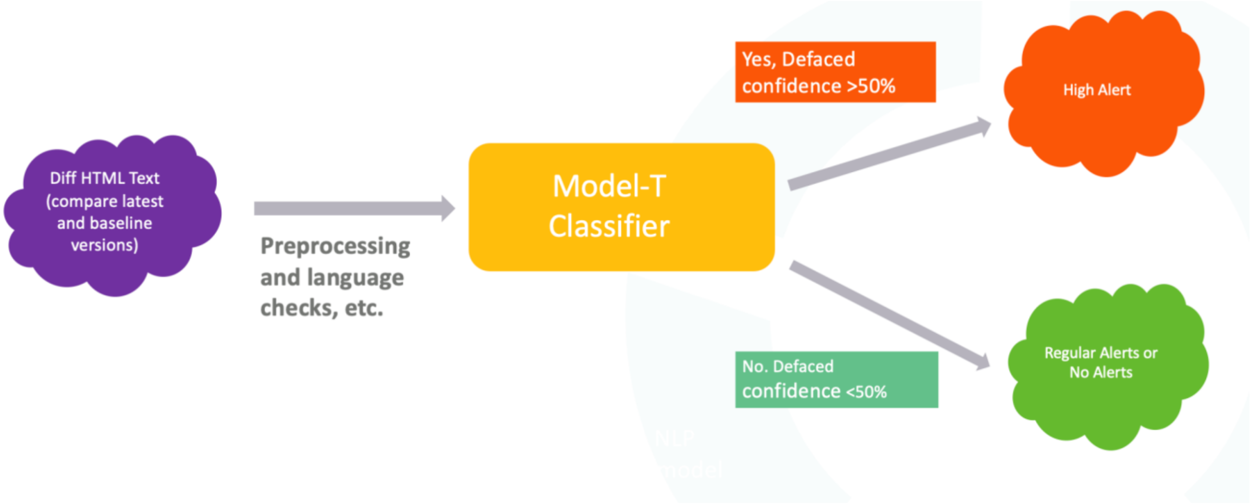
The new NLP Engine utilizes two state- of-the-art NLP classification models, the Model T and Model X.

Model T is a transformer-based model which adopts the mechanism of self-attention that simulates human comprehension when reading web pages. It takes the webpage text content without the stop words (by, what, that, etc.) as input. It then analyzes the tokens(words) and extracts semantic information to understand the text. Like human beings, it differentially weights the significance of each part of the input data, analyzing the context to extract the key information. Eventually, this produces a defacement confidence score to represent the engine's analysis about whether the text is part of a web defacement attack.

Based on the considerations mentioned above, an AI NLP model has been optimized to provide a classification score based on any HTML text submissions. Internal testing reveals an accuracy rate of over 90% for the real-world HTML text that is supplied to the model.


Past 10 years of defaced data
Non-defaced data
Model Training
Further testing with Present Data
Final Optimal Model-T Classifier
Transfer Learning
Retraining and Optimization
Well-known NLP Transformer model

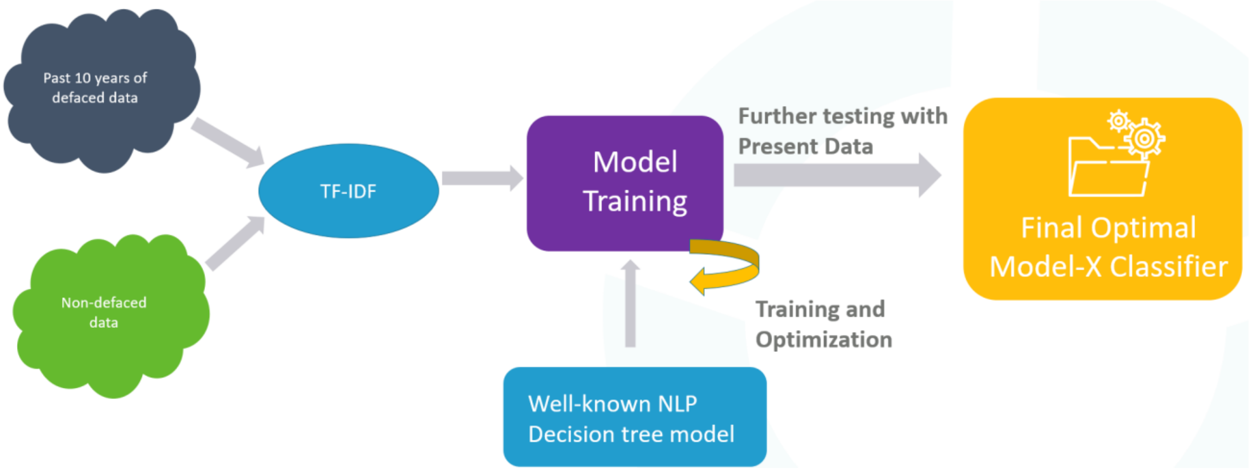
Once the final optimal Model-T Classifier has been trained, Model-T is integrated with the WebOrion monitoring platform so it can perform inference each time it checks a web page.


Diff HTML Text (compare latest and baseline versions)
Preprocessing and language checks, etc.
Model-T Classifier
Yes, Defaced confidence >50%
High Alert
No. Defaced confidence <50%
Regular Alerts or No Alerts
NLP
model

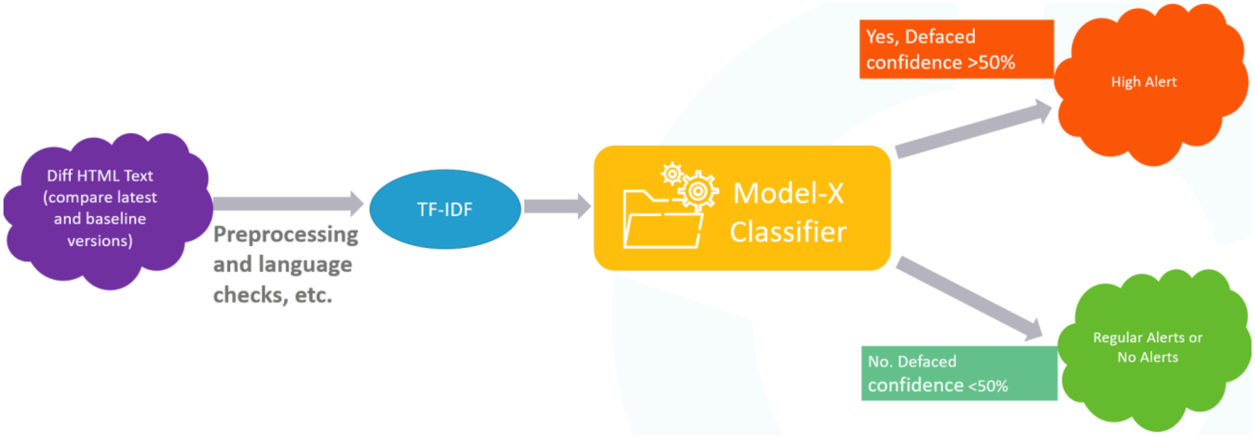
In real world production systems, WebOrion will check for changes on the monitored webpage on a regular basis. Any HTML text change will then be pre-processed (e.g., language checks, removal of stop words, etc.) and sent to Model-T classifier. Model-T will then perform its AI NLP analysis and provides a defacement classification (yes/no) and confidence level (e.g., 80%). Webpage changes that are classified as 'defaced' will get a high alert while non-defaced text will get regular or no alerts. In this manner, WebOrion customers can prioritize response to high alerts, differentiate between regular alerts and reducing false positives .

Model X Training and Inference

The Model X is developed to understand and analyze the text content in a different angle. The Model X is a well-known decision-tree-based model that has often outperformed other algorithms or frameworks. It takes the tf-idf vectorized text as input, which counts the occurrence of a word in one sample and then down-weighs it with the occurrence of the same word over all the samples. This is because normally the text in the defaced web page is significantly different from that of non-defaced web pages. By analyzing the word frequencies occurring in the text content, Model X learns to differentiate between defaced and non-defaced web pages from these extracted statistics. Performance tests have demonstrated that the Model X can achieve around 90% accuracy on real word HTML data.


Past 10 years of defaced data
Non-defaced data
TF-IDF
Model Training
Further testing with Present Data
Final Optimal Model-X Classifier
Training and Optimization
Well-known NLP Decision tree model

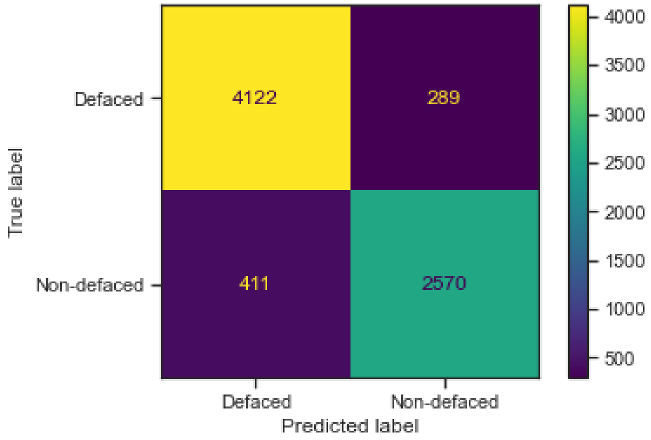
After the final Model-X Classifier has been optimized, Model-X is integrated with the WebOrion monitoring platform to incorporate with Model-T to verify the web pages.


Diff HTML Text (compare latest and baseline versions)
Preprocessing and language checks, etc.
TF-IDF
Model-X Classifier
Yes, Defaced confidence >50%
High Alert
No. Defaced confidence <50%
Regular Alerts or No Alerts

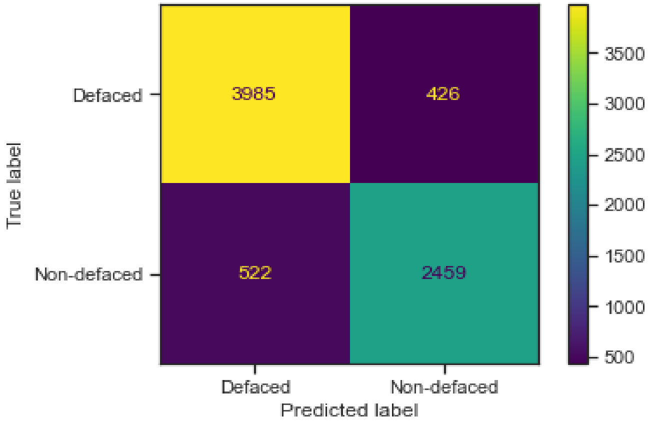
During actual verification, when changes are detected, the HTML text runs through a similar pre-process as Model X and is sent to the Model X classifier. Model X will analyze the words frequency information and give a defacement classification (yes/no) and confidence level (e.g. 80%) like Model-T. The Model T and Model X make predictions independently and the alert will be sent out.

Model Accuracies

The AI Models often require massive quantities of training data to learn. Fortunately, taking advantage of the datasets with defaced web pages collected over the past 10 years. Both models are trained on defaced and non-defaced datasets with more than 120 thousand samples. The Model T and Model X have each learned the patterns of the defaced web pages over the past 10 years. Both models were then tested using a dataset of 4000 defaced webpages that had not been run through before. Combining the prediction of both of the models, an accuracy of 95% to identify all the defaced web pages was achieved.

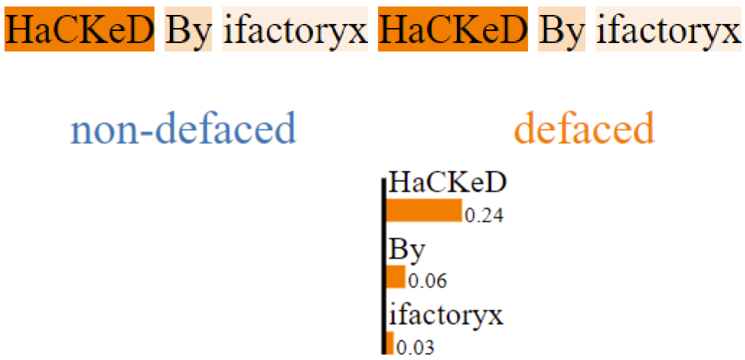


Confusion Matrix for Model T



Confusion Matrix for Model X

One of the effective methods to understand the models' predictions is the interpretability of the model. A series of experiments were done to understand the model interpretability. An example is provided below:



HaCKeD By ifactoryx HaCKeD By ifactoryx

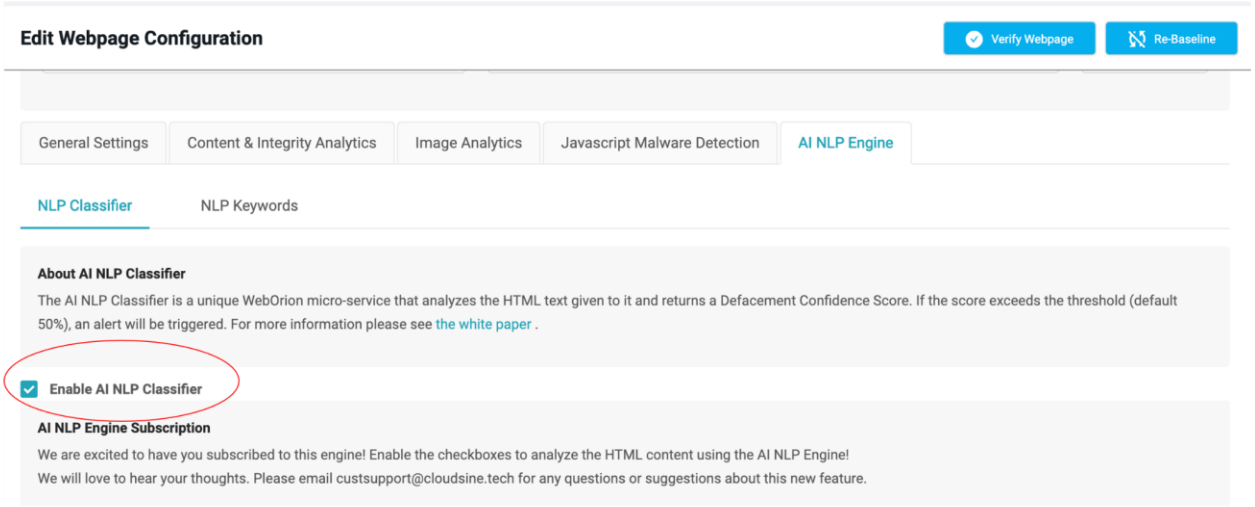
The model captured the word "HaCKeD," normally detected on the defaced page, and then gives a high score for this word. Since the model knows the organization name "ifactoryx," the word also influences the model prediction to determine with greater confidence that a defacement has occurred.

How to Use and Configure AI NLP

The new AI NLP Engine focuses on analyzing the text contents inside a webpage to determine whether the webpage is defaced or not. We use two state-of-the-art AI NLP classifier models, Model T and Model X. Both models have been trained on the defaced and non-defaced datasets collected over the past 10 years. Users that are new to the AI NLP system are recommended to first familiarize with implementing the Model-T Classifier for their website(s).

Here are some of the key steps to enable the AI NLP Engine:

Step 1: Enable the AI NLP Engine:



About AI NLP Classifier

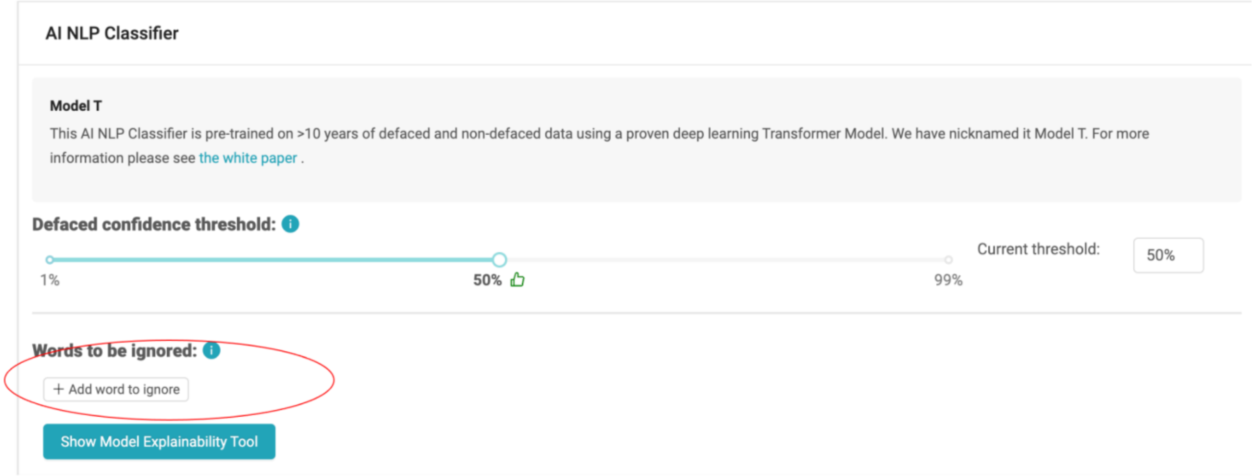
The AI NLP Classifier is a unique WebOrion micro-service that analyzes the HTML text given to it and returns a Defacement Confidence Score. If the score exceeds the threshold (default 50%), an alert will be triggered. For more information please see the white paper .

☑ Enable AI NLP Classifier

AI NLP Engine Subscription

We are excited to have you subscribed to this engine! Enable the checkboxes to analyze the HTML content using the AI NLP Engine!
We will love to hear your thoughts. Please email custsupport@cloudsine.tech for any questions or suggestions about this new feature.

Step 2: Add Any Words to be Ignored by the AI NLP Classifier (E.g., Company Name)



Model T

This AI NLP Classifier is pre-trained on >10 years of defaced and non-defaced data using a proven deep learning Transformer Model. We have nicknamed it Model T. For more information please see the white paper .

Defaced confidence threshold: 50%

Current threshold: 50%

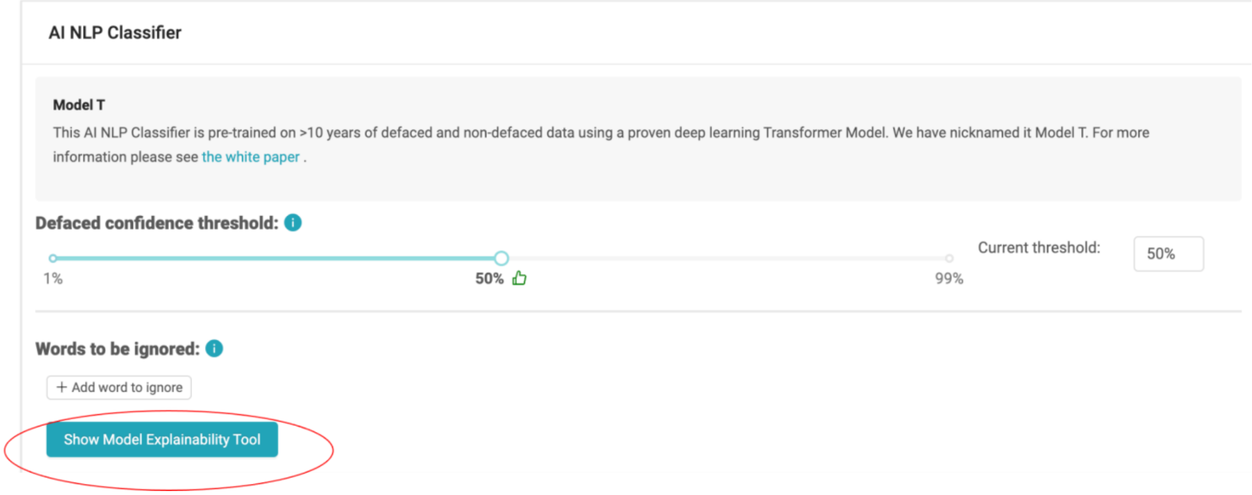
Words to be ignored:

Note: The defacement confidence threshold is set to 50% by default, which has been optimized during the AI NLP training phase. If the defacement confidence exceeds this threshold, that means the model considers the webpage as defaced. A high alert notification will then be sent out to the designated persons.

It is not recommended to set the threshold to the extremes ends of the 1 to 99% meter . A threshold that is set too high a threshold may miss some actual alerts, while a very low threshold will result in more false alerts.

Step 3: Show Model Explainability Tool

After enabling the AI NLP model, users can click the "Show Model Explainability Tool" which will be covered in the next section.



AI NLP Classifier

Model T

This AI NLP Classifier is pre-trained on >10 years of defaced and non-defaced data using a proven deep learning Transformer Model. We have nicknamed it Model T. For more information please see the white paper .

Defaced confidence threshold:

Current threshold: 50%

Words to be ignored:

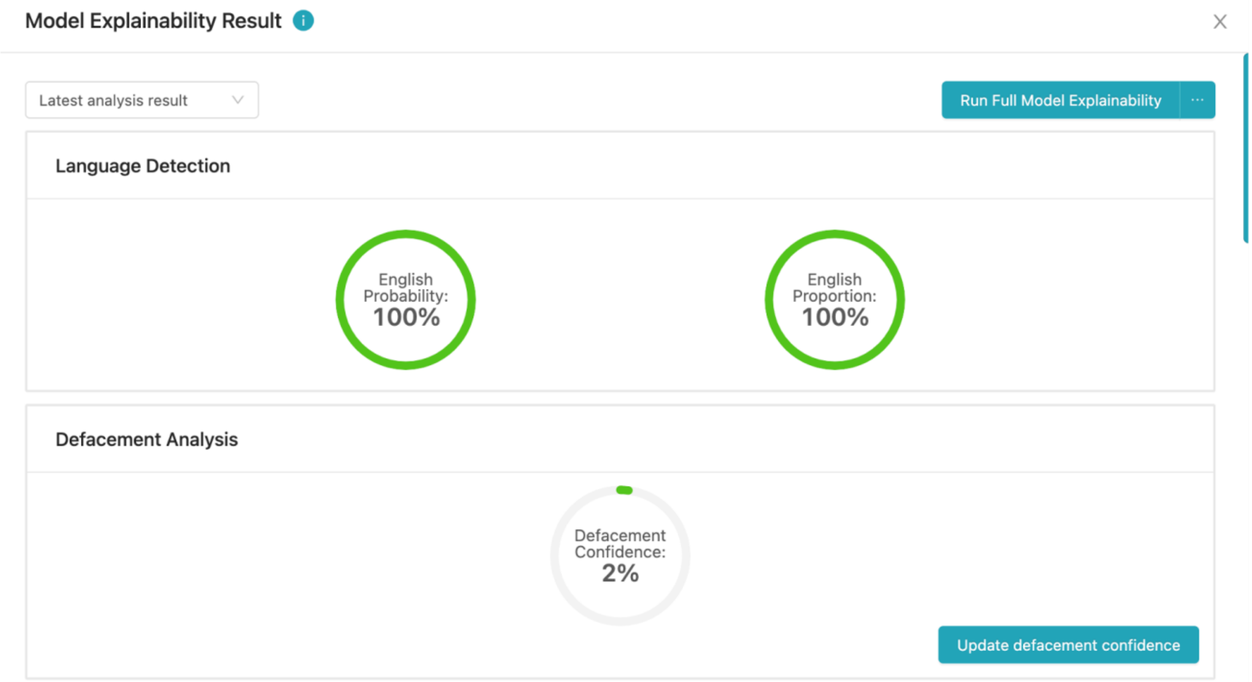
Explainability Tool

Beyond just providing an AI Blackbox which is provided the classification results, an AI NLP Explainability Tool has been added to decipher the results of the AI NLP engine.

In a typical scenario, the AI NLP Engine may send out alerts when normal updates are detected. This is because the AI NLP Engine might not have learned the words such as company or product names. Thus, the engine may classify the webpage as defaced when first encountering these words.

WebOrion engineers have made the deliberate effort to include this tool to explain these prediction results. The tool provide additional insights into why the AI NLP engine considers the webpage as defaced. This is done by assigning a weight to each word, which is an indication of how much that word leads the model to make such a prediction. Any words with a positive weight add to the confidence level that the webpage is defaced. Words that are assigned the highest weights will be ignored in for future predictions to improve classification accuracy for the monitored webpage.

Step 4: Check Explainability Tool



Model Explainability Result

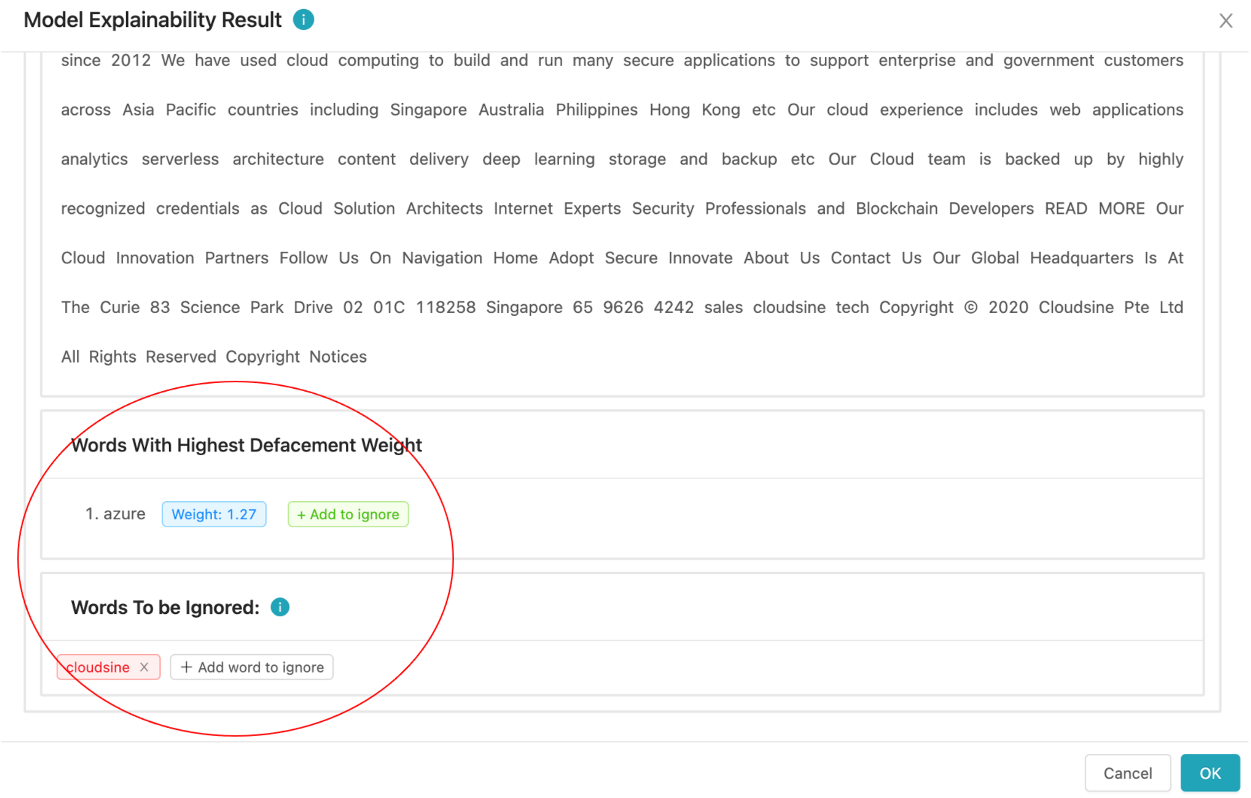
Language Detection

English Probability: 100%

English Proportion: 100%

Defacement Analysis

Defacement Confidence: 2%



Model Explainability Result

since 2012 We have used cloud computing to build and run many secure applications to support enterprise and government customers across Asia Pacific countries including Singapore Australia Philippines Hong Kong etc Our cloud experience includes web applications analytics serverless architecture content delivery deep learning storage and backup etc Our Cloud team is backed up by highly recognized credentials as Cloud Solution Architects Internet Experts Security Professionals and Blockchain Developers READ MORE Our Cloud Innovation Partners Follow Us On Navigation Home Adopt Secure Innovate About Us Contact Us Our Global Headquarters Is At The Curie 83 Science Park Drive 02 01C 118258 Singapore 65 9626 4242 sales cloudsine tech Copyright © 2020 Cloudsine Pte Ltd All Rights Reserved Copyright Notices

Words With Highest Defacement Weight

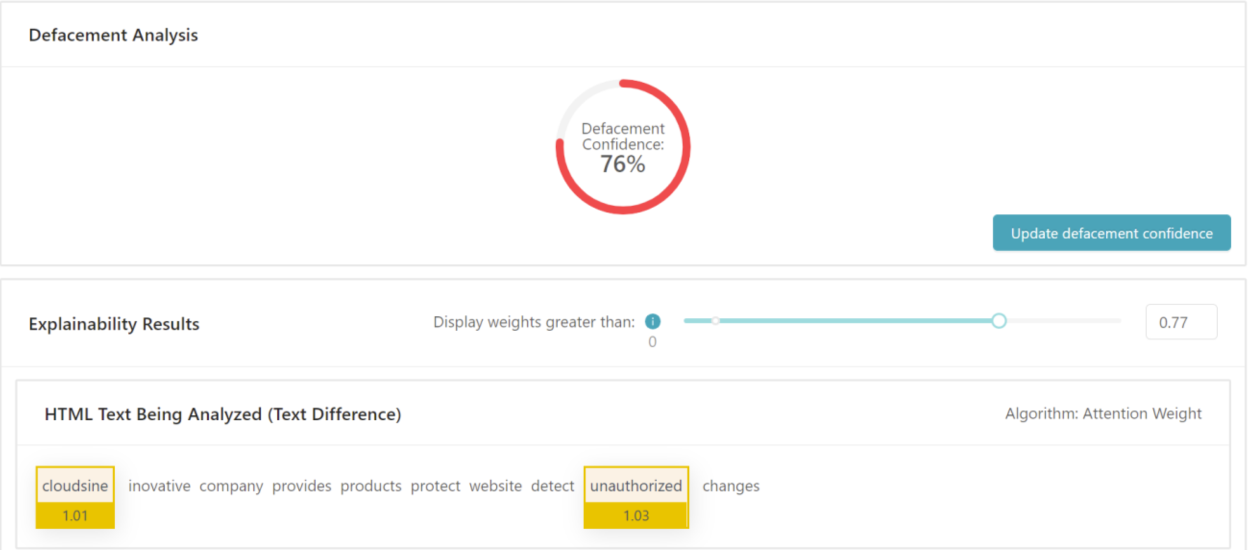
1. azure Weight: 1.27

Words To be Ignored:

cloudsine ×

Step 5: Tuning the Engine to Adapt to the Website

In AI NLP, it is impossible for the model to know all English words and sentences as the language evolves over time. A common misclassification could be due to company names (e.g., Cloudsine). To address this issue, we provide AI NLP interpretability tests on the web page to find out which words cause the model to make wrong predictions. After finding out these words, the user can easily assign them to an ignore list. In future predictions, these words will be ignored in the Classifier, to reduce the occurrence of false positives.



Defacement Analysis

Defacement Confidence: 76%

Explainability Results

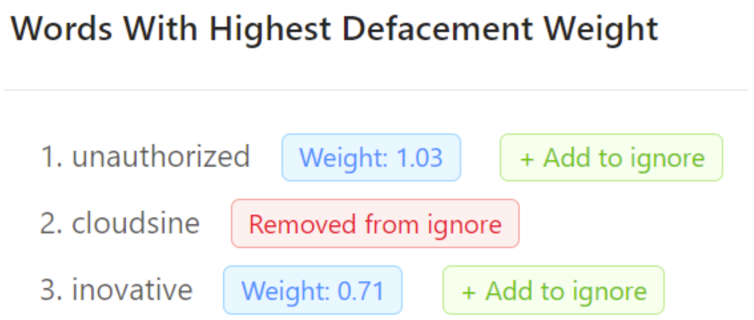
Display weights greater than: 0.77

HTML Text Being Analyzed (Text Difference)

Algorithm: Attention Weight

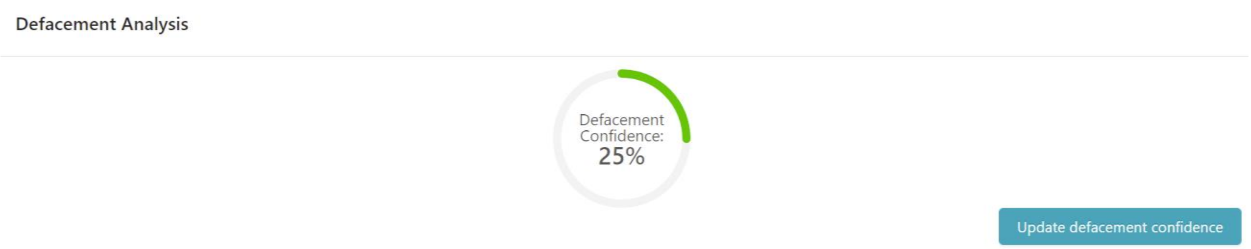
cloudsine 1.01 inovative company provides products protect website detect unauthorized 1.03 changes

As shown in the example above, the company name "cloudsine" is something the model has not learned before and it contributes significantly to the defacement confidence score. The word "cloudsine" can be ignored after and further re-runs result in a defacement confidence score that is significantly lower. This is an example of how AI NLP can be tuned for different customer websites.



Words With Highest Defacement Weight

1. unauthorized Weight: 1.03
2. cloudsine Removed from ignore
3. inovative Weight: 0.71

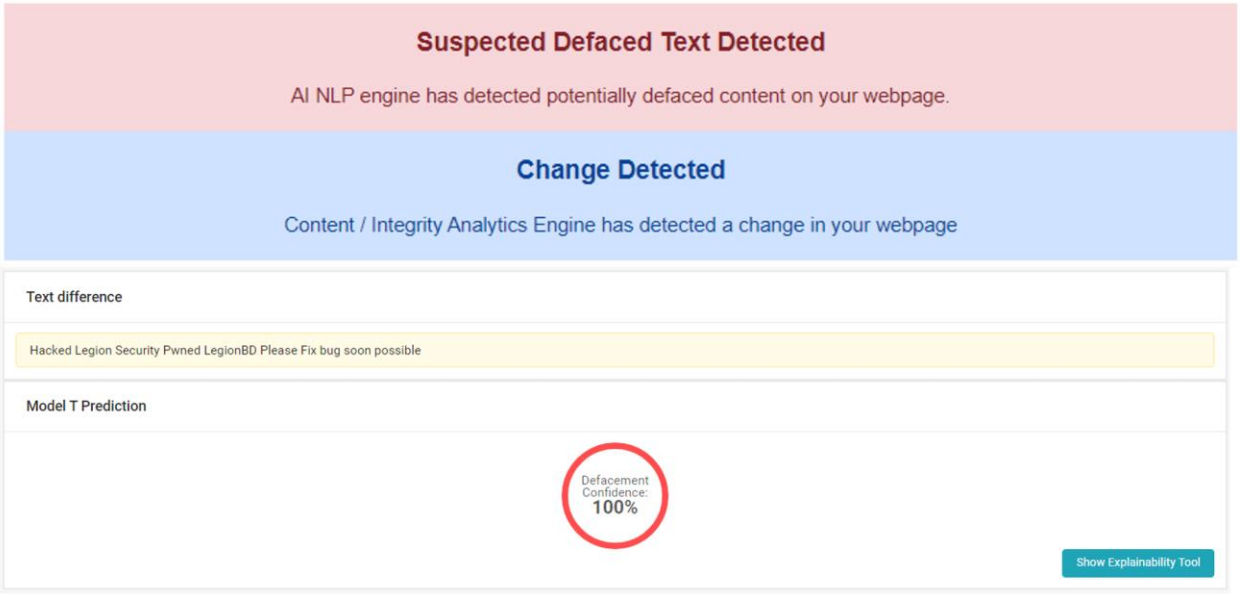


Defacement Analysis

Defacement Confidence: 25%

Step 6: Observe Alerts and Finetune if Required

Email alerts will also be highlighted with red color text if AI NLP has classified something as defaced. The sample is shown below.



Suspected Defaced Text Detected

AI NLP engine has detected potentially defaced content on your webpage.

Change Detected

Content / Integrity Analytics Engine has detected a change in your webpage

Text difference

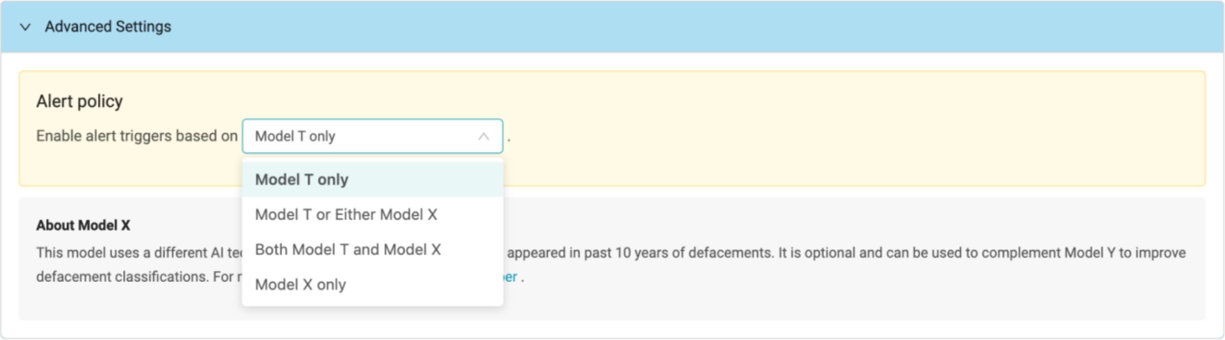
Hacked Legion Security Pwned LegionBD Please Fix bug soon possible

Model T Prediction

Defacement Confidence: 100%

Advanced Configurations with Multiple Models

Model-T and Model-X can be used as standalone models to classify HTML text for web defacements. For more advanced users, both models can be combined to provide greater AI model robustness and flexibility to monitor various kinds of websites. The diagram below shows the configuration options to select one of (1) Model T only (default), (2) Either Model T or Model X, (3) Both Model T and Model X and (4) Model X only.



Advanced Settings

Alert policy

Enable alert triggers based on Model T only .

About Model X

This model uses a different AI te appeared in past 10 years of defacements. It is optional and can be used to complement Model Y to improve defacement classifications. For r er .

Both models hosted on the AI NLP Engine can be run in "Either" or "Both" modes. "Either" means high alerts will be triggered if one of Model-T or Model-X detects a >51% defacement confidence. This mode may lead to increased false positives from the AI NLP engine but will provide the most stringent monitoring mode. As such, "" can be selected when customers are on high alert for any potential cyber crisis impacting their websites.

The "Both" mode means that high alerts will only be triggered if BOTH Model-T and Model-X classify the text as defaced text with >51% confidence. This mode will reduce false positives but with the minor risk of missing actual defaced text.

The table below summarizes the various modes and scenarios. The Model-T is recommended to be implemented first prior to the addition of other modes as required.

	Scenario 1: Model-T >51%, Model-X >51%	*Scenario 2: Model-T >51%, Model-X <49%*	*Scenario 3: Model-T <49%, Model-X >51%*	*Scenario 4: Model-T <49%, Model-X <49%*
Model-T only	High Alert	High Alert	Regular Alerts	Regular Alerts
Model-X only	High Alert	Regular Alerts	High Alert	Regular Alerts
Either Model	High Alert	High Alert	High Alert	Regular Alerts
Both Models	High Alert	Regular Alerts	Regular Alerts	Regular Alerts

Why AI NLP is better than Change and Keyword Detection.

The WebOrion Monitor's Content Integrity Engine already detects any webpage changes and notifies the user when a change is detected. Although this mechanism indicates the changes made (E.g., a JavaScript or HTML change), it does not know whether the changes are malicious or not. With the AI NLP Engine, it will analyze the changes before sending out alerts. If the model considers the webpage as defaced, it will send out a high alert with warnings of a defacement to users. But if the model does not consider the webpage as defaced, regular alerts will be sent to inform the users that some changes have been detected.

Dictionary words that contain typical hacker language and names simply detect the presence of these words in the web page text contents. The challenge with this method is that it cannot detect new hacker names that are not in the black-listed keyword dictionaries. The AI NLP Engine analyzes the context and semantics when making predictions, thus making it much more intelligent than keyword detection. This also allows AI NLP Engine to detect defacements by new hacker groups without explicitly learning about their names.

Conclusions of Using AI NLP Classifier

Using the AI NLP Classifier provides the following benefits:

- Beyond change or keyword detection, using the AI NLP Classifier adds a new level of intelligence to analyze and classify the type of changes on monitored webpages.
- Detection of the language used in webpages.
- The Explainability tool helps users to better understand the reasoning behind the AI engine classification of webpage contents.
- Provides differentiated alerts for high-risk changes.
- Reduces false positives if the system is configured to send no alerts for HTML text changes that are not classified as defaced.

The launch of the AI NLP engine complements the WebOrion Monitor's other advanced engines to make it the most comprehensive offering in the industry. The AI NLP engine is easily available in the WebOrion Monitor AI Pack. Do reach out to us via our website if you wish to know more!