

Fuse Censor Extension

Extension de censure intelligente de mots sur toutes vos pages Internet !

Pour les navigateurs :



Firefox



Chrome



Edge

Charlier Fabien
Debouchage Antoine
Golbol Danyel
Ollivier Pierre
Sajus Alexandre

Lien vers le dépôt Gitlab :
<https://gitlab-cw2.centralesupelec.fr/fabien.charlier/swear-analyzer>



Description du produit final

→ Pourquoi ?

- Se protéger des mauvais contenus présents un peu partout sur le net.
- Embellir son expérience utilisateur

→ Pour qui ?

Personnes sensibles, enfants, personnes âgées...
Tous les proches que vous voulez protéger !



Utilisation simple et intuitive accessible à tous

MVP : Masquer les mots et expressions à caractère offensant pour l'utilisateur

→ Mode spoiler

- Censure rapide et efficace des mots à caractère violent ou insultant
- Système d'affichage des mots injurieux, avec votre consentement

→ Mode Gentleman

- Remplacement automatique et intelligent de ces mots pour une lecture pacifiste et agréable

Modes

Spoiler Mode



I am a little funny text and I love cheesecake. Thank you for your understanding



I am a [REDACTED] and I [REDACTED] cheesecake. [REDACTED] for your understanding

I am a **little funny text** and I **love** cheesecake. **Thank you** for your understanding

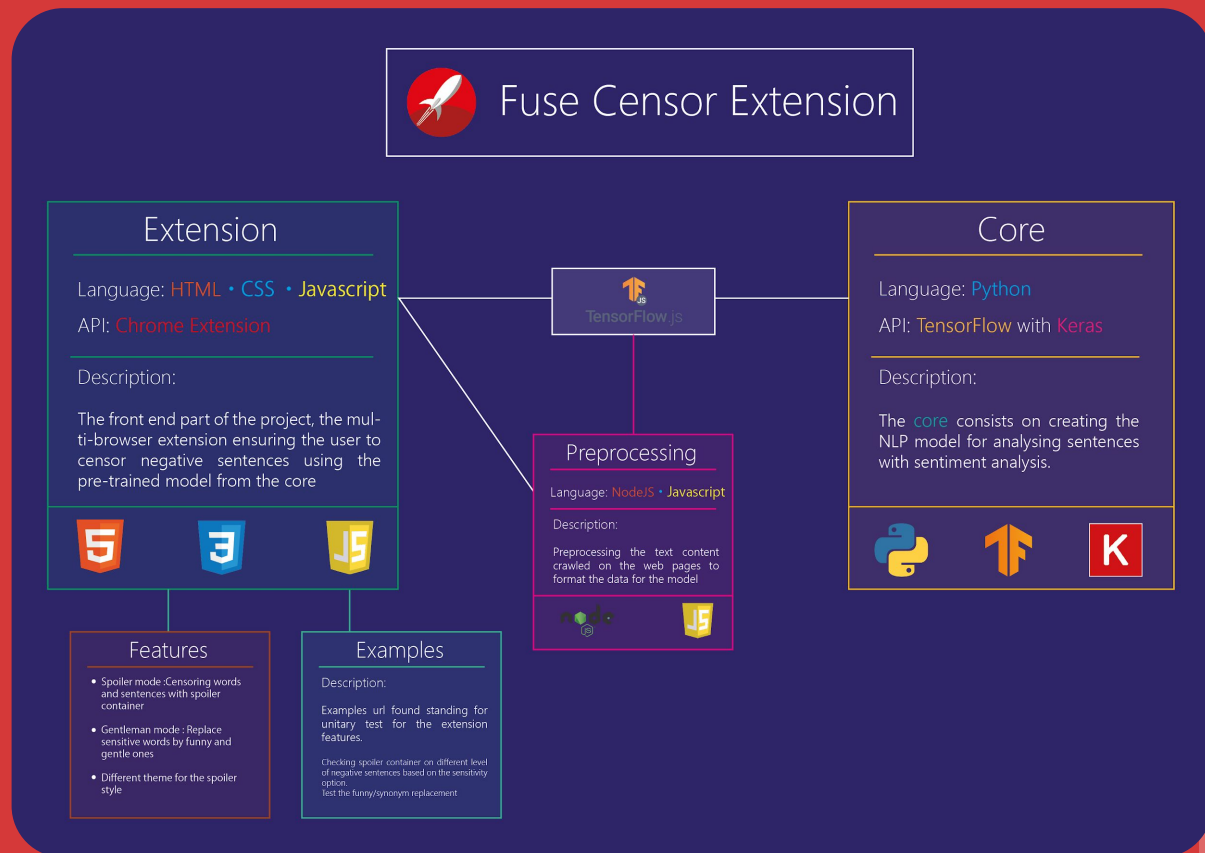
Gentleman Mode



You bastard. You vulgar little maggot. You worthless bag of filth.

You beautiful person. You vulgar little chips. You worthless bag of flower.

Structuration du projet





Fuse Censor is an extension that provides an efficient way to censor swear words and negative contents based on sentiment analysis through Natural Language Processing.

Enable



Options

Language



English

Mode

Gentleman

Theme

Bright

Sensibility



Obscenity



Threat



Apply

Découpage

Extension (html, css, javascript)

Danyel ⇒ réalisation du popup

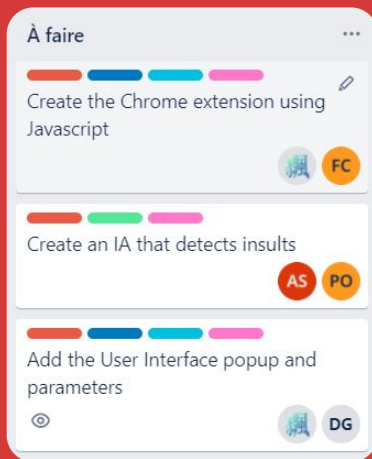
Antoine ⇒ création de l'extension

Intelligence artificielle (Python)

Alexandre ⇒ Détection des mots injurieux

Pierre ⇒ Détection des mots injurieux

Répartition des tâches sur Trello



Fonctionnalités

Fabien ⇒ Gentleman & Spoiler mode

Antoine ⇒ Spoiler mode

Danyel ⇒ Thèmes

Le popup (Html, Javascript & CSS)

```
<td>Theme</td>
<td style=" left:51 ; position:relative">
  <div style="text-indent: 15px;">
    <select name="style" id="theme" width="100px">
      <option value="dark">Dark</option>
      <option value="bright">Bright</option>
      <option value="red">Red</option>
    </select>
  </div>
</td>
```


```
span.spoiler-default.dark {
  background-color: #51, 51, 51;
  color: #51, 51, 51;
}
```

```
span.spoiler-default.dark:hover {
  background-color: #85, 85, 85;
  color: #85, 85, 85;
}
```

```
span.spoiler-active.dark {
  background-color: #85, 85, 85;
  color: white;
}
```

```
.slider {
  position: absolute;
  cursor: pointer;
  top: 0;
  left: 0;
  right: 0;
  bottom: 5;
  background-color: #191919;
  border: 1px solid #aaa;
  border-radius: 30px;
  transition: all 0.4s;
}
/* The sliding button */
.slider:before {
  position: absolute;
  content: "";
  width: 15px;
  height: 15px;
  left: 3px;
  top: 1px;
  background-color: #fff4f4;
  border-radius: 18px;
  transition: all 0.4s;
}
/* On checked */
input:checked + .slider {
  background-color: #444a5f;
  border-color: #0062be;
}
input:checked + .slider:before {
```

```
span.addEventListener('click', () => {
  if (span.className === "spoiler-default" + ' ' + settings.theme) {
    span.className = "spoiler-active" + ' ' + settings.theme;
  }
  else if (span.className === "spoiler-active" + ' ' + settings.theme) {
    span.className = "spoiler-default" + ' ' + settings.theme;
  }
});
this.element.appendChild(span);
```

 **Fuse Censor Extension** 

Fuse Censor is an extension that provides an efficient way to censor swear words and negative contents based on sentiment analysis through Natural Language Processing.

Enable ☒

Options

Language English

Mode Gentleman

Theme Dark

Sensibility Strong High Medium Low Weak

Obscenity Strong High Medium Low Weak

Threat Strong High Medium Low Weak

Apply

Design et feuille de style (CSS)

I am a little funny text and I love cheesecake. Thank you for your understanding

I am a [REDACTED] and I [REDACTED] cheesecake. [REDACTED] for your understanding

I am a little funny text and I love cheesecake. Thank you for your understanding

```
class Slider {  
  /**  
   * Some style  
   */  
  static thumbStyle = `  
    width: 18px;  
    height: 1px;  
    margin: -8px 0 0;  
    border-radius: 50%;  
    cursor: pointer;  
    border: 0 !important;  
  `;  
  static prefs = ['webkit-slider-runnable-track', 'moz-range-track', 'ms-track'];  
  
  /**  
   * SlideButton constructor  
   * @param classname The class name of the template  
   * @param color Specify the color of the active part of the slider  
   */  
  constructor(classname, color) {  
    this.classname = classname;  
    this.container = $(`<div class="${this.classname}"></div>`);  
    this.input = this.container.find('input');  
    this.labels = this.container.find('label');  
    this.its = this.labels.find('li');  
    this.slide = this.container.find('div.active-slide');  
  
    this.color = (color != null) ? color : '#37adb8';  
  
    this.value = $(this.input)[0].value;  
  
    this.sheet = document.createElement('style');  
    document.body.appendChild(this.sheet);  
  }  
}
```



Fuse Censor Extension



Fuse Censor is an extension that provides an efficient way to censor swear words and negative contents based on sentiment analysis through Natural Language Processing.

Enable



Options

Language

English

Mode

Gentleman

Theme

Dark

Sensibility



Obscenity



Threat



Apply

Détection de phrases

```
▼<div id="mw-content-text" lang="fr" dir="ltr" class="mw-content-ltr">
  ▼<div class="mw-parser-output">
    ▶<div class="bandeau-container homonymie plainlinks">...</div>
      <p class="mw-empty-elt">

    </p>
    ▶<table class="infobox_v2">...</table>
  ..
  ▶<p>...</p> == $0
  ▶<p>...</p>
  ▼<p>
    "L'."
    <a href="/wiki/Anglais" title="Anglais">anglais</a>
    " a été la première langue utilisée, et "
    <a href="/wiki/Wikip%C3%A9dia_en_anglais" title="Wikipédia en anglais">Wikipédia en anglais</a>
    " compte plus de six millions d'articles début "
    <time>2020</time>
    ". "
    <a href="/wiki/Wikip%C3%A9dia_en_fran%C3%A7ais" title="Wikipédia en français">Wikipédia en français</a>
    ", ouverte le "
    <time class="nowrap" datetime="2001-03-23" data-sort-value="2001-03-23">23 mars 2001</time>
    ", compte un peu plus de deux millions d'articles la même année. Wikipédia existe en plus de 300 langues, dans une apparence unie, mais avec de grandes variations de contenus.
    "
  </p>
```

```
▼<div class="mw-page-container">
  <a class="mw-jump-link" href="#content">Aller
  ▼<div class="mw-page-container-inner">
    <input type="checkbox" id="mw-sidebar-check
    checkbox" checked>
    ▶<header class="mw-header">...</header>
  ..
  ▼<div class="mw-workspace-container"> == $0
    ▶<div id="mw-panel" class="mw-sidebar">...</div>
    ▶<div id="mw-navigation">...</div>
    ▼<div class="mw-content-container">
      <!-- Please do not use role attribute a
      -->
      ▼<main id="content" class="mw-body" role=
        <a id="top"></a>
        ▶<div id="siteNotice" class="mw-body-co
        ▼<div class="mw-indicators mw-body-cont
          ▶<div id="mw-indicator-protection-edi
          </div>
          ▼<div id="mw-indicator-protection-rer
            ▼<div class="nopopups">
              ▼<a href="//fr.wikipedia.org/wiki/
              Page_au_nom_prot%C3%A9g%C3%A9" tit
              peut être modifié.">
              <img alt="Le titre de cette pa
              upload.wikimedia.org/wikipedia
```


Remplacement de mots

You bastard. You vulgar little maggot. You worthless bag of filth.

CENSORED



Non

You beautiful person.
You vulgar little chips.
You worthless bag of flower.

Oui

Remplacement

```
const VB = ['let', 'see', 'eat', 'play', 'move', 'cast', 'chose'];
const VBD = ['ate', 'flowered', 'powered', 'cost', 'did', 'fell'];
const VBG = ['playing', 'teaching', 'forgetting', 'forgot', 'paid'];
const VBN = ['eaten', 'built', 'cut', 'quit', 'run', 'shaken', 'spread'];
const VBP = ['let', 'see', 'eat', 'play', 'move', 'cast', 'chose'];
const VBZ = ['eats', 'does', 'plays', 'pushes', 'moves', 'tells', 'wakes'];
const NN = ['beautiful person', 'fish', 'table', 'sheep', 'chips', 'water', 'flower', 'gentleman'];
const NNS = ['dogs', 'gentlemen', 'feminists', 'cows', 'bottles', 'mouses', 'beds'];
const NNP = ['London', 'Smith', 'Paris', 'James'];
const NNPS = ['Smiths', 'Jones', 'Browns'];
const JJ = ['good', 'fast', 'late', 'happy', 'young', 'smart', 'brave', 'noble', 'wise'];
const FW = ['rendez-vous', 'café', 'faux pas', 'persona non grata', 'adieu', 'baguette'];
const JJR = ['better', 'faster', 'happier', 'younger', 'smarter', 'braver', 'nobler', 'wiser', 'taller', 'farther'];
const JJS = ['best', 'fastest', 'latest', 'happiest', 'youngest', 'smartest', 'bravest', 'noblest', 'wisest', 'dearest'];
```

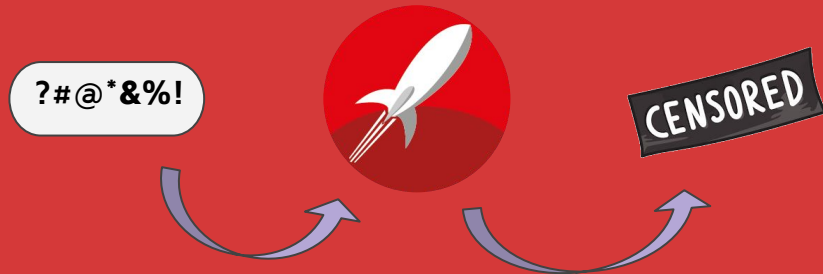
Cœur de l'extension

→ Objectifs

- Analyser chaque phrase d'un texte
- Pour chacune, donner dans quelle mesure elle est obscène, menaçante, toxique

→ Moyens

Utilisation d'une intelligence artificielle permettant de décortiquer chaque phrase



→ Simple

- Conversion facile vers JS
- Léger
- Rapide

→ Entrée

Un texte

→ Sortie

Une évaluation du texte selon plusieurs critères

- Toxicité
- Obscénité
- Menace

Cœur de l'extension

`["toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate"]`

Phrase 0

...

Phrase 4

0.6947695	0.06126317	0.4801368	0.08376214	0.4075508	0.07023317
0.7122015	0.05662242	0.48156792	0.08738527	0.4168801	0.08322236
0.7984909	0.13696933	0.82305884	0.09854552	0.6013024	0.18301272
0.11448562	0.01472524	0.037099	0.00970423	0.02614838	0.00813499
0.8014021	0.13386497	0.82005084	0.09691969	0.5969834	0.17772505

Pre-processing

→ Dataset: Kaggle Toxic Comment Classification Challenge

	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
Explanation	\r\nwhy the edits made under my use...	0	0	0	0	0	0
	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
	"\r\nMore\r\nI can't make any real suggestions...	0	0	0	0	0	0
	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

→ Lowercase:

```
why the edits made under my username hardcore metallica fan were reverted?
```

→ Tokenize:

```
[76, 1, 136, 119, 177, 29, 557, 4636, 1221, 83, 328, 53, 2334,
```

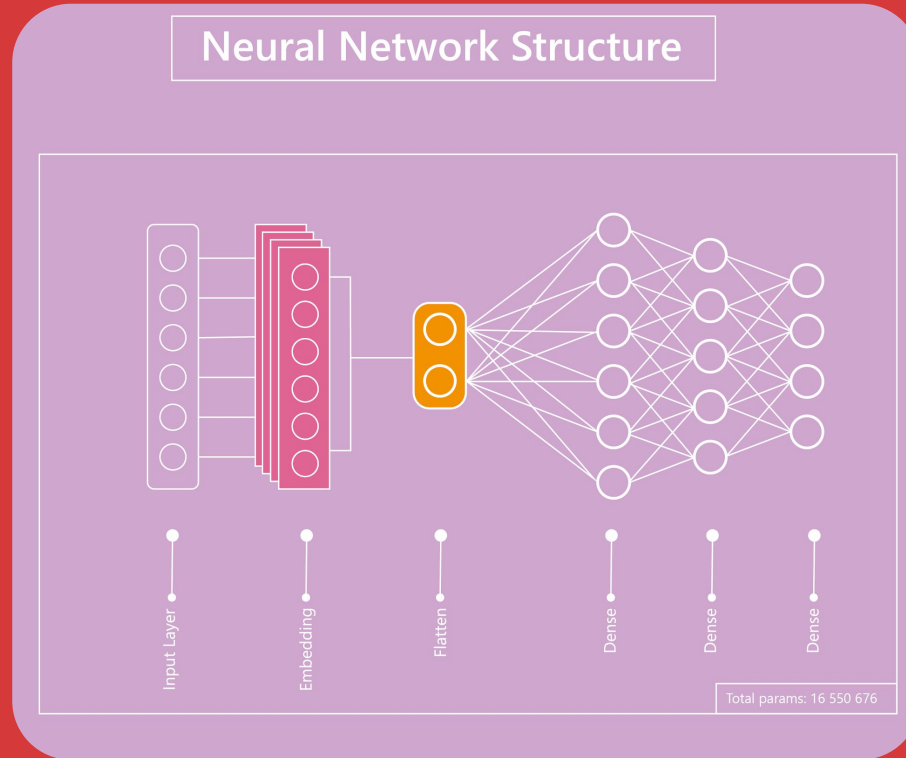
→ Padding:

```
[ 76   1  136  119  177  29  557  
 16   64 4998  147   7 3851   35  
  1  404   33   1   40   28  143  
  0   0   0   0   0   0   0  
  0   0   0   0   0   0   0
```

→ Tokenizer:

```
{'the': 1, '\r': 2, 'to': 3, 'of': 4, 'and': 5,  
 'if': 23, 'was': 24, 'or': 25, 'article': 26,  
 42, 'has': 43, 'all': 44, 'no': 45, 'will': 46,  
 1, 'should': 62, 'here': 63, 'some': 64, 'see':  
 80, 'use': 81, 'only': 82, 'were': 83, 'when': 8  
 m': 99, 'balls': 100, 'make': 101, 'good': 102,  
 way': 116, 'such': 117, 'sources': 118, 'made':  
 32, 'pages': 133, 'need': 134, 'section': 135,
```

Les différentes couches utilisées



Les différentes couches utilisées

→ Embedding

- Phrases n'ayant pas la même longueur
- Solution : redimensionner les phrases, ignorer les 0 (masque)

I	am	a	little	funny	text	and	I	love	cheesecake
Thank	you	for	your	understanding					



1	2	3	4	5	6	7	1	8	9
10	11	12	13	14	0	0	0	0	0

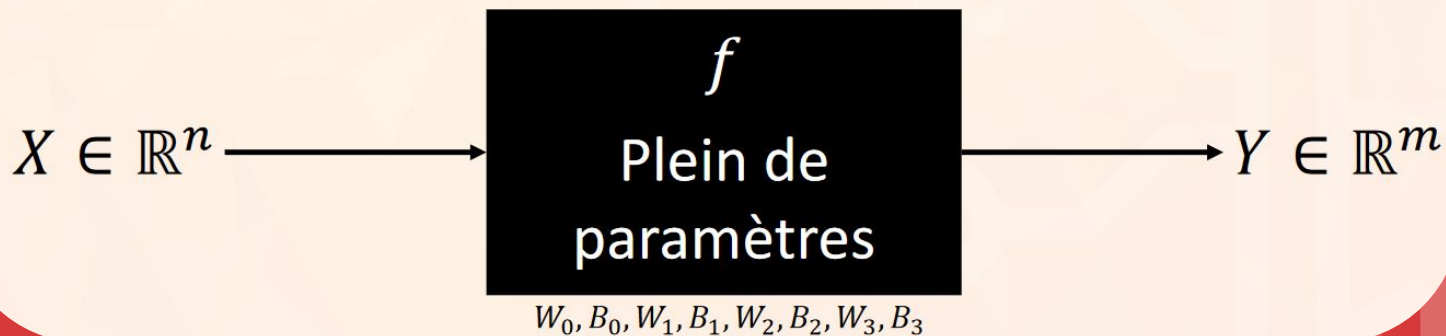
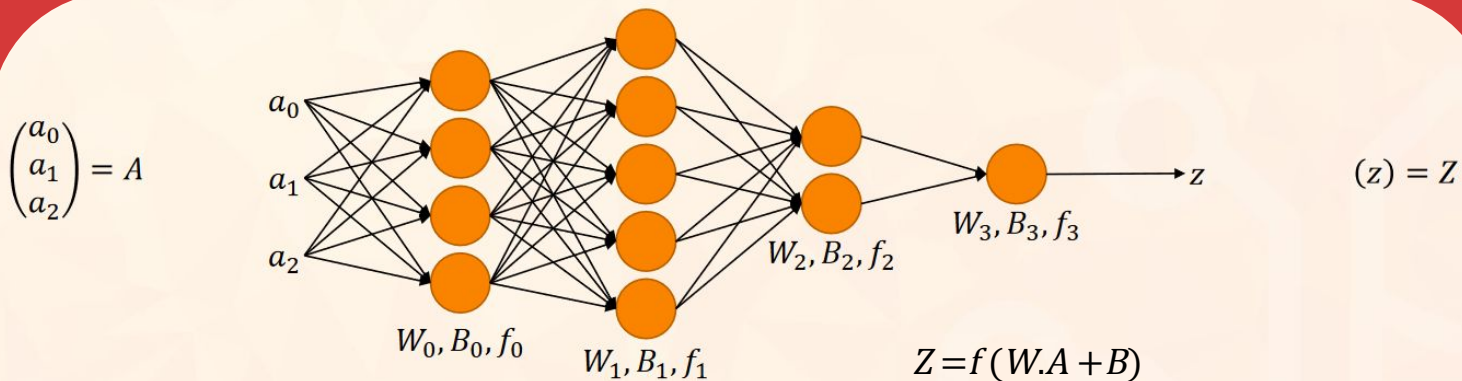
→ Flatten

Permet de passer à une dimension (aplatissement)

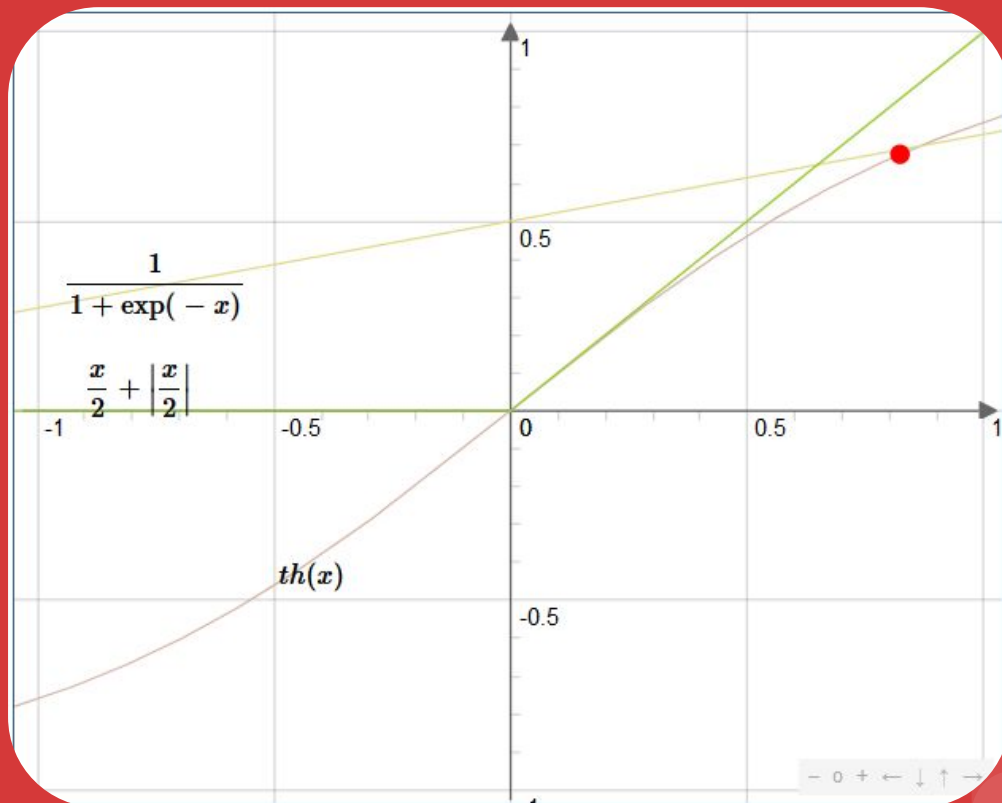
→ Dense

- Passer progressivement de beaucoup de neurones à 6 neurones
- Fonctions d'activation associés à des poids (paramètres) pour séparer les valeurs élevées des valeurs faibles

Couches denses



Fonctions d'activation



Sortie finale

```
["toxic", "severe_toxic", "obscene", "threat", "insult", "identity_hate"]
```

Phrase 0

...

Phrase 4

```
[[0.6947695 0.06126317 0.4801368 0.08376214 0.4075508 0.07023317]  
 [0.7122015 0.05662242 0.48156792 0.08738527 0.4168801 0.08322236]  
 [0.7984909 0.13696933 0.82305884 0.09854552 0.6013024 0.18301272]  
 [0.11448562 0.01472524 0.037099 0.00970423 0.02614838 0.00813499]  
 [0.8014021 0.13386497 0.82005084 0.09691969 0.5969834 0.17772505]]
```

CENSORED

Critiques du modèle

- Modèle d'IA non suffisant par rapport à nos attentes
- Outils d'analyse et d'enregistrement plus poussés
- Seule la langue anglaise est prise en compte
- Choix de balises restreint (temps)

