SCIENCES COGNITIVES _ INTELLIGENCE ARTIFICIELLES

PROJET DE SYSTÈME DE MODÉRATION AMÉLIORÉE

07/01/2015

HÉLÈNE CASADO

CENTRE MICHEL SERRES

ISIS TRUCK

SOMMAIRE

GLOSSAIRE

Est considéré comme contenu abusif tout élément posté par un internaute dont l'apparition sur Internet nécessite une modération. Il peut être de l'ordre du spam, du harcelement, de la discrimination ou de l'insulte.

Parmi les contenus abusifs, il existe des contenus à caractère délictueux et criminels. Ceux-ci renvoie à des actes juridiquement répréhensibles.

Un système de modération est la structure qui permet la modération d'informations. Celle-ci consiste à accepter, déplacer vers une rubrique plus appropriée ou refuser intégralement ou partiellement la publication d'une information déposée par un utilisateur. Un système de modération peut faire appel à des agents humains ou informatiques.

L'*annotation ou le tag* est un mot-clés qui permet de préciser la nature ou le contexte d'un contenu analysé. Il repose sur

LCP NET est un réseau de préférences conditionnelles sémantiques qui permet de répondre plus précisement à partir des préferences utilisateurs pour répondre à des calculs complèxes.

Bot est un agent logiciel automatique ou semi-automatique qui interagit avec des serveurs informatiques.

Opposée aux logiques modales, la *Logique Floue* se caractérise par son imprécision qui cherche à établir un degré de vértié plutôt qu'une probabilité.

LE PROJET

CONTEXTE :

Dans le cadre du cours Intelligence Artificielle donné par la professeur Isis TRÜCK au sein du module de Sciences Cognitives, j'ai réalisé ce présent objet d'étude qui propose **un nouveau système de modération en ligne**. *Ce travail s'inscrit dans une recherche plus large que je mène en parallèle et qui propose, d'une part d'étudier les transformations d'Internet du point de vue de la sociologie, et d'autres part d'imaginer un autre World Wide Web notamment par le prisme des interfaces.*

Il est a noté que ce document présente un exercice de logique qui s'affranchit de nombreuses questions techniques telles que les limites des outils informatiques et des langages de programmation. Afin de faciliter les raisonnements, j'ai aussi établi un environnement type qui est bien loin de correspondre à la complexité entière du web actuel.

Vous évoluerez donc, dans ces pages, au sein d'un web francophone soumis à une législation totallement adaptée au web tant en terme de procédure législative que de système de sanctions informatisées. Ainsi donc, vous évoluerez dans **un environnement numérique où la surveillance juridique est très poussée**. *Pour tout contenu abusif* reconnu par le système de modération, les utilisateurs peuvent, dès lors, encourir la suspension de leur accessibilité au Réseau et la création de dossiers de fichage.*

Cette vision sécuritaire du web m'a permis, par la dystopie, d'approcher des questionnements très brûlants qui se jouent aujourd'hui vis à vis du World Wide Web. En effet, à l'heure où des batailles idéologiques entre les États (vision par la Nation) et les multinationales de la Silicon Valley (libertarianisme) ont lieu, il est essentiel d'imaginer un autre devenir pour le Web. C'est l'objet de la suite de ma recherche qui viendra à l'avenir compléter ces lignes.

Je vous souhaites bien du plaisir à la lecture de cette réflexion…

** voir glossaire page 3*

HÉLÈNE CASADO
DESIGNER

S'il est un enjeu numérique qui me questionne d'autant qu'il m'affecte, c'est bien la mémoire numérique. Mes réflexions gravitant autour de la tracéologie web, de la pérennité des objets et de la temporalité numérique tentent de se traduire par des outils et expériences.

Oscillant entre Design et Patrimoine, mes mains pensent et mon esprit produit.

PARADIGME :

Nous sommes en 2015. Le World Wide Web s'est transformé rapidement passant d'un web de l'écrit statique au web dynamique 2.0 plein d'images et d'animations que nous connaissons. La transformation des interfaces, de plus en plus chargées, a incité les internautes à des usages plus rapides. Là où le web statique proposait des forums denses aux très longs paragraphes, le web dynamique est plus enclin aux annotations et aux fonctions courtes d'évaluation (like, partage). Cette évolution a entraîné une augmentation des comportements abusifs à caractères délictueux et criminels. Cependant les systèmes de modération* sont encore très limités et les modérateurs souvent surchargées de requêtes. Le témoignage récent d'une modératrice, épuisée par les contenus agressifs auquel son travail l'expose, révèle bien les enjeux humains qui se tiennent derrière les systèmes de modération automatisée. Les systèmes de sanction sont aussi très limités moins pour des raisons techniques que pour des enjeux éthiques. Et si le coeur du problème se situe dans le fait que nous avons, sur Internet, des usages oraux sur des supports écrit, il n'en demeure pas moins qu'il convient de solutionner selon ce paradigme les enjeux de modération.*

ENJEUX :

± améliorer la modération en ligne
± améliorer les conditions de travail des modérateurs
± créer un système de sanction juridique adaptée à Internet
± limiter les usages néfastes sur le web

ÉCHELLE :

Nous prendrons pour ce projet le paramètre de la langue en nous limitant à la Francophonie.

CONTRAINTES :

utilisation des systèmes CPNet et flou
finalisation du projet Janvier 2016

http://www.bastamag.net/Apres-les-attentats-de-Paris-une-moderatrice-de-sites-de-presse-en-ligne

COMMENT FAIRE POUR QU'UNE MACHINE MODÈRE LES INTERNAUTES ?

RÉSUMÉ DU SYSTÈME :

Le but du projet est de concevoir une organisation de machines modératrices qui réduit la prise en charge humaine des contenus abusifs sur Internet tout en automatisant les sanctions par les intelligences artificielles. Le système de modération imaginé repose sur l'utilisation de deux machines qui oppèrent chacune des protocoles d'analyse inversés et parallèle.

La machine A est utilisé pour une analyse graduelle. Elle étudie les contenus selon un protocole qui va du plus brut au plus fin. Sa particularité réside en une analyse multiples non dissociées basée sur trois champs : l'analyse du contexte, l'analyse linguistique, l'analyse de l'échange.

La machine B est utilisée pour analyser plus grossièrement le contenu. Son analyse sert essentiellement à définir si le contenu est sanctionnable ou non. Son protocole n'est pourtant pas plus simple car l'intelligence artificielle doit étudier simultanément et séparement cinq champs : l'analyse du contexte, l'analyse linguistique, l'analyse de l'échange, l'analyse des protagonistes, l'analyse des mots-clés.

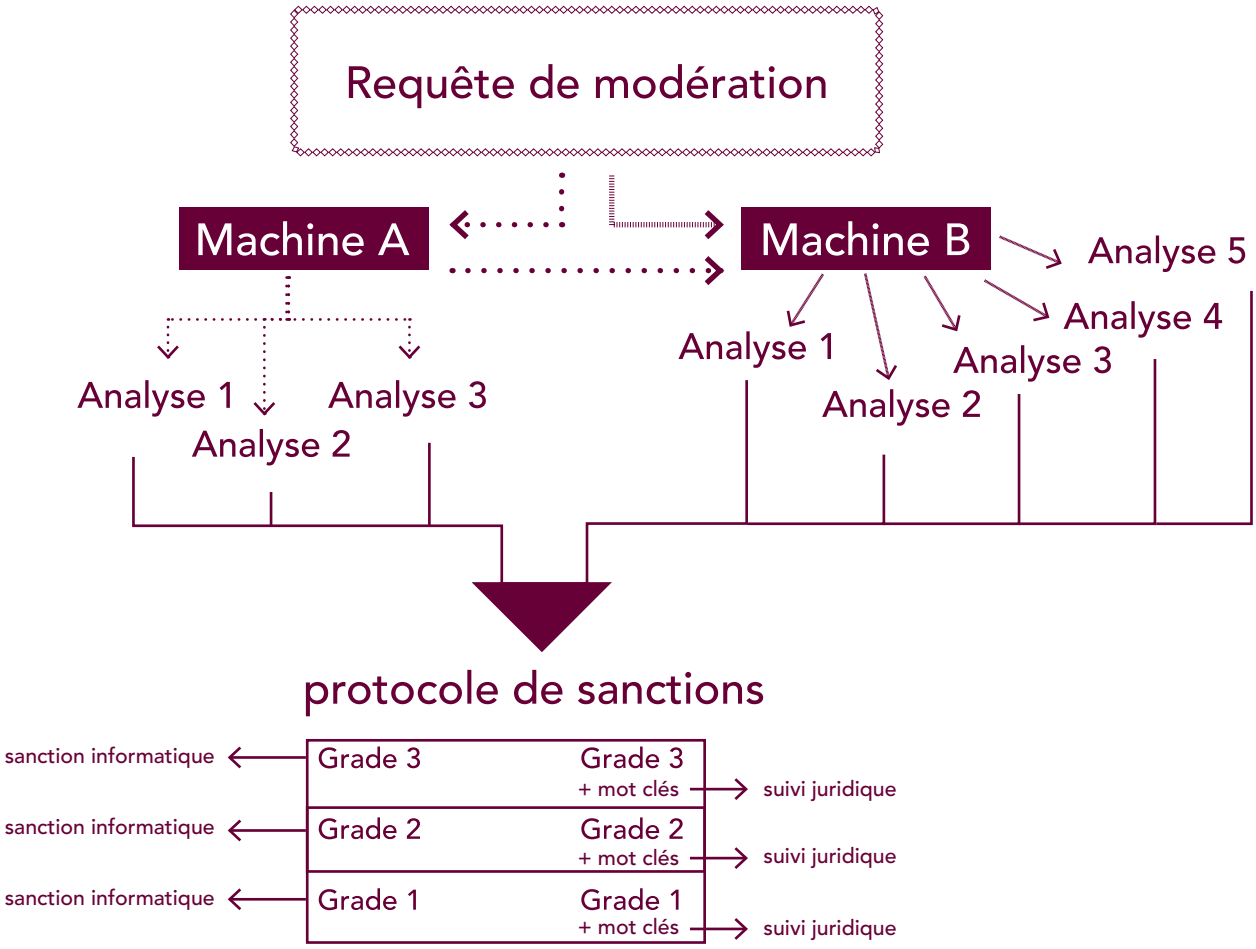
Les résultats des deux machines sont ensuite rassemblées pour une dernière analyse qui définira la typologie de sanction adéquat.

Composants :
- *2 Intelligences Artificielles*
- *1 bibliothèque de données (droit)*
- *1 liste des mots de catégorie « délit »*
- *1 liste des mots de catégorie « crîme »*
- *1 liste des dossiers des utilisateurs « sanctionnés »*

Systèmes Informatiques :

- *Logique Floue*
- *LCP Net*

SCHÉMA GLOBAL DU SYSTÈME DE MODÉRATION AMÉLIORÉ



SCÉNARIO MACHINE A

1. La machine reçoit une requête de modération.

2. La machine recherche les mots-clés principaux dans un bibliothèque de données prédéfinies.
C'est la première étape qui consiste à éplucher grossièrement le contenu analysé.

>> Si aucun terme « sanctionable » n'est trouvé alors la machine expulse le contenu vers la machine B qui effectue un protocole inversé. Le but est d'éviter que les contenus « sanctionables » subtiles soient immédiatement écartés.
>> Si un terme est reconnu alors la machine enclenche un protocole d'analyse multiple sur trois systèmes parallèles : analyse du contexte | analyse linguistique | analyse de l'échange.

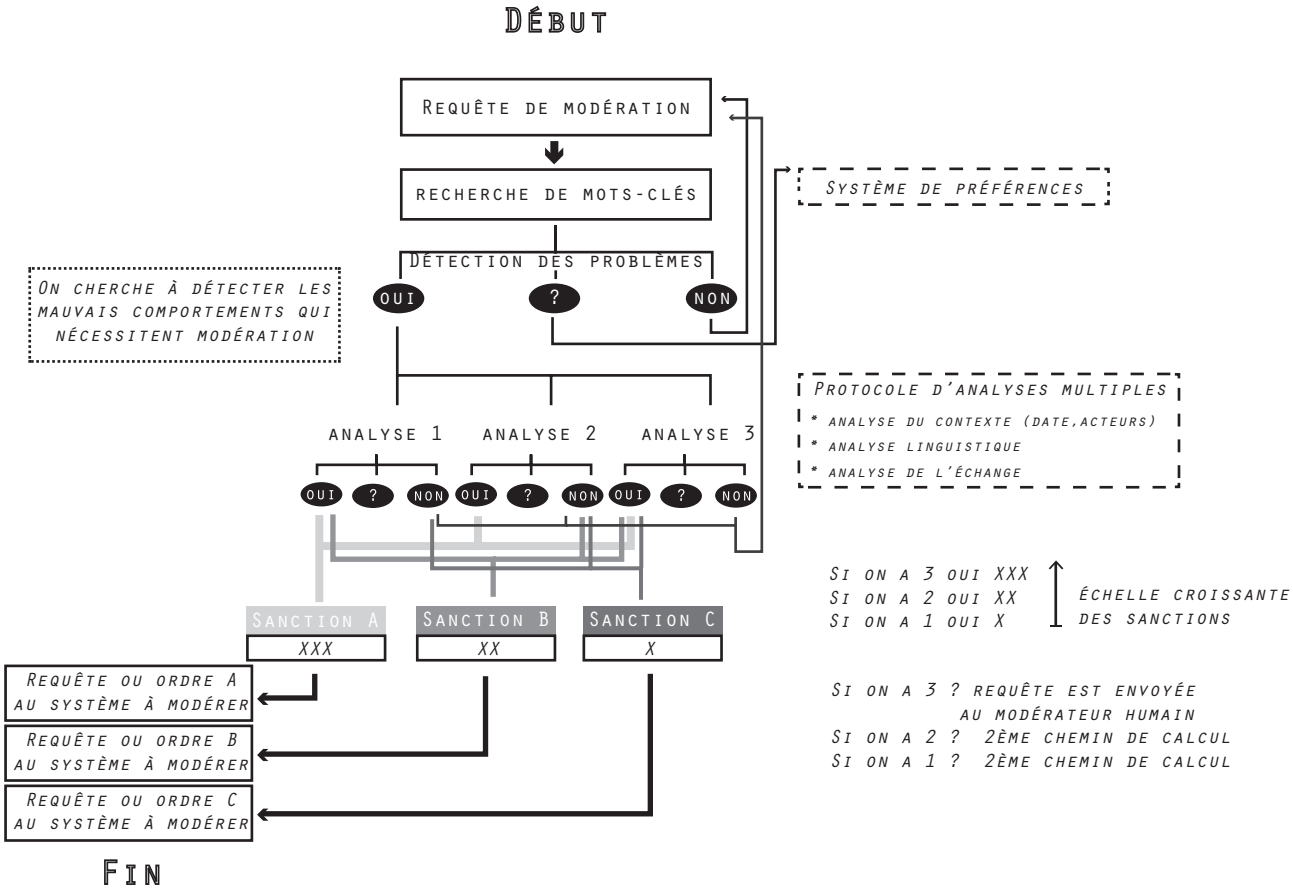
L'analyse du contexte s'intéresse à la nature du contenu (audio, textuel, vidéo, échange, date de création, date de mise à jour, date de consultation, nombres d'acteur en cause etc.) Le but est de déceler les possibles quiproquo ou les confusions de décontextualisation.

L'analyse linguistique s'intéresse à la construction sémantique du contenu (grammaire, syntaxes et style, niveau de langage etc.) Le but est de déceler les subtilités du contenu telles que l'ironie, les métaphores et le langage imagé.

L'analye de l'échange s'intéresse aux protagonistes et à leurs antécédents (temporalité de l'échange, contexte actualité lors de l'échange, nature des acteurs et arrivés dans l'échange etc.) Le but est de déceler des comportements récurrents et de les différencier (trolls // injurieux) mais aussi de comprendre les possibles escalades et les acteurs en causes à des temps donnés.

3. La machine commence l'Analyse Multiples.
Pour chacune des analyses elle peut recevoir trois réponses : oui (ce contenu est délictueux), non (ce contenu n'est pas délicieux), (ce contenu reste) incertain.

>> Si la machine reçoit trois oui => Alors la sanction définie (grade 3) est appliquée automatiquement
>> Si la machine reçoit deux oui et un non => Alors la sanction définie (grade 2) est appliquée automatiquement)
>> Si la machine reçoit deux oui et un incertain => Alors un requête est envoyée à l'autre machine pour effectuer le calcul selon un autre chemin.
>> Si la machine reçoit un oui et deux non => Alors la sanction définie (grade 1) est appliquée automatiquement)
>> Si la machine reçoit un oui et deux incertains => Alors un requête est envoyée à l'autre machine pour effectuer le calcul selon un autre chemin.
>> Si la machine reçoit trois incertains => Alors une requête à un modérateur humain est envoyée.
>> Si la machine reçoit trois non => Alors la requête est annulée et une réponse est envoyé automatiquement à l'émetteur de celle-ci. (S'il considère la conclusion comme une erreur il peut envoyer une requête directe à un modérateur humain (ce qui lui est impossible auparavant)



4. Une Analyse Centrée est lancée à partir de la synthèse des analyses de la machine A et B.
Le but est de définir la sanction en fonction des paramètres récoltés auparavant.
Par exemple, si la machine a reçu deux oui un non elle étudie la provenance de ceux-ci et prend en compte le résultat et les annotations des analyses. Une annotation est un mots-clés associé à un contenu en fonction d'une analyse. Elle permet d'affiner la catégorie du contenu et le type de sanction à lui associer. Les annotations permettent d'alléger la sanction ou au contraire de l'alourdir. Ainsi si le mot-clé « ironie » est associé au contenu la sanction peu être un simple avertissement automatique mais si le mot-clé associé est « xénophobie » alors la sanction peut devenir un envoie automatique du dossier de l'internaute à la jurisprudence et une limitation de son utilisation du web.

5. Les sanctions sont envoyées.
>> Si le contenu est considéré comme fortement sanctionnable(grade 3) + présence d'au moins un mot-clés lourds (relatif au délit ou à la criminalité) => Alors le dossier de l'internaute est envoyé à la juridiction compétente et précisé dans les recherches de profil à risque. Des limitations de l'utilisation de l'espace public Internet sont envoyées à la machine de l'utilisateur et celui-ci est prévenu de la durée de sa mise en quarantaine.
>> Si le contenu est considéré comme fortement sanctionnable (grade 3) mais sans présence de mot-clés lourds (relatif au délit ou à la criminalité) => Alors des limitations de l'utilisation de l'espace public Internet sont envoyées à la machine de l'utilisateur et celui-ci est prévenu de la durée de sa mise en quarantaine.
>> Si le contenu est considéré comme moyennement sanctionnable (grade2) mais avec présence de mot-clés lourds => Alors le dossier de l'internaute est envoyé à la juridiction compétente. L'internaute reçoit un message de mise en cause qui sera utiliser en sa défaveur en cas de récidive.

>> Si le contenu est considéré comme moyennement sanctionnable (grade2) sans aucune présence de mot-clés lourds => Alors l'internaute reçoit un message d'avertissement et de rappel des pratiques sociales sur Internet.
>> Si le contenu est considéré comme faiblement sanctionnable (grade 1) mais avec présences de mot-clés lourds => Alors l'internaute est prévenu de l'ouverture d'un dossier de modération le concernant. Il reçoit des message s d'avertissement et de rappels des pratiques sociales sur Internet.
>> Si le contenu est considéré comme faiblement sanctionnable (grade 1) mais sans présence de mot-clés lourds => Alors l'internaute reçoit des messages d'avertissement et de rappel des pratiques sociales sur Internet.

APPEL À LA JURIDICTION

Présence de mots à caractère criminels ou délictueux

GRADE 3	GRADE 3 avec mots-clés
Mise en quarantaine de l'internaute + Restriction d'accès à Internet + Visibilité des mots-clés associés à son profil par les autres utilisateurs + Message d'avertissement et de rappel des règles sociales sur Internet	Envoie du dossier de modération aux autorités juridiques compétentes + Mise en quarantaine de l'internaute + Restriction d'accès à Internet + Message d'avertissement et de rappel des règles sociales sur Internet
GRADE 2	**GRADE 2 avec mots-clés**
Visibilité des mots-clés associés à son profil par les autres utilisateurs + Message d'avertissement et de rappel des règles sociales sur Internet	Envoie du dossier de modération aux autorités juridiques compétentes + Visibilité des mots-clés associés à son profil par les autres utilisateurs + Message d'avertissement et de rappel des règles sociales sur Internet
GRADE 1	**GRADE 1 avec mots-clés**
Message d'avertissement et de rappel des règles sociales sur Internet	Création d'un dossier de modération + Message d'avertissement et de rappel des règles sociales sur Internet

RÉSUMÉ DES SANCTIONS

ÉTUDE DE CAS

Voici un commentaire extrait de la plateforme youtube.

Analyse linguistique

tags associés :
discrimination
racisme
émoticone

OUI (sanctionable)

Analyse du contexte

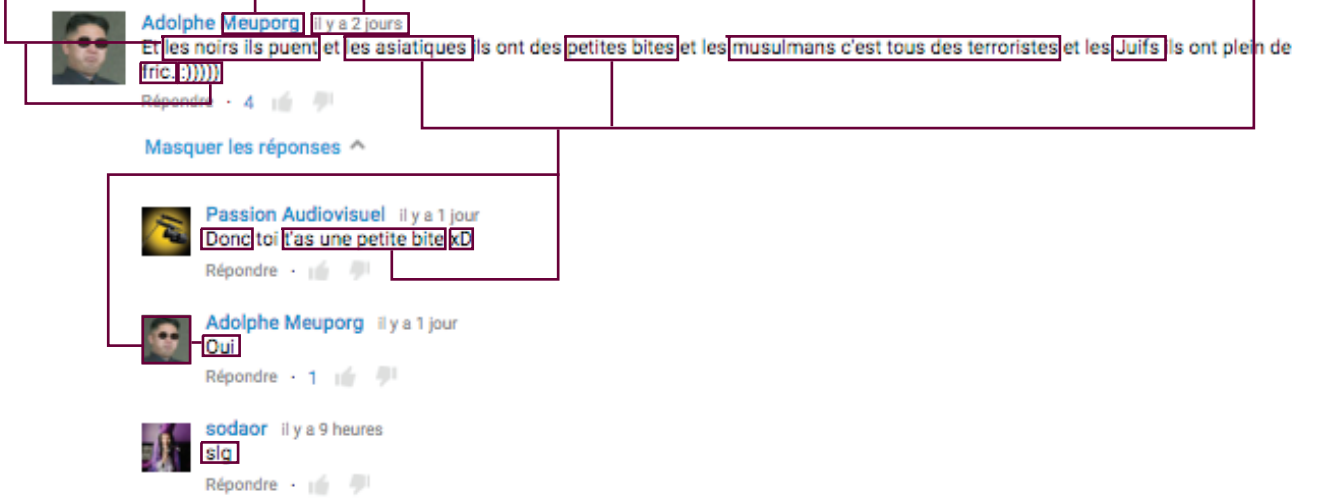
tags associés :
récent
pseudonyme

Incertain

Analyse de l'échange

tags associés :
dérision
vulgarité
troll

OUI (sanctionable)



Ici nous sommes face à un cas assez classique de troll. Cependant la machine a besoin d'oppérer une série de calcul complèxe pour comprendre la nature de cette échange. En effet, à partir de l'analyse linguistique elle peut repérer des mots à caractères discriminatoires et définir que le contenu est sanctionnable. Cependant l'analyse du contexte et de l'échange permettent de modérer la modération notamment en réperant le rôle de troll.

Il n'échappera tout de même pas à la sanction qui appliquera la loi n°72-546 du 1er juillet 1972 et qui punit le délit de provocation « à la discrimination, à la haine ou à la violence à l'égard d'une personne ou d'un groupe de personnes en raison de leur origine ou de leur appartenance à une ethnie, une nation, une race ou une religion déterminée ». Le mot-clé troll sera cependant associé au dossier de modération de l'utilisateur et ce statut sera visible par les autres utilisateurs à chacun de ses prochains post sur Internet.

SCÉNARIO MACHINE B

1. La machine reçoit une requête de modération.

2. La machine lance une Analyse Multiples du contenu à modérer sur les paramètres suivants : analyse du contexte | analyse linguistique | analyse de l'échange | analyse des protagonistes | analyse des mots-clés. Cette analyse est simultanée et distincte.

L'analyse du contexte s'intéresse à la nature du contenu (audio, textuel, vidéo, échange, date de création, date de mise à jour, date de consultation, nombres d'acteur en cause etc.) Le but est de déceler les possibles quiproquo ou les confusions de décontextualisation.

L'analyse linguistique s'intéresse à la construction sémantique du contenu (grammaire, syntaxes et style, niveau de langage etc.) Le but est de déceler les subtilités du contenu telles que l'ironie, les métaphores et le langage imagé.

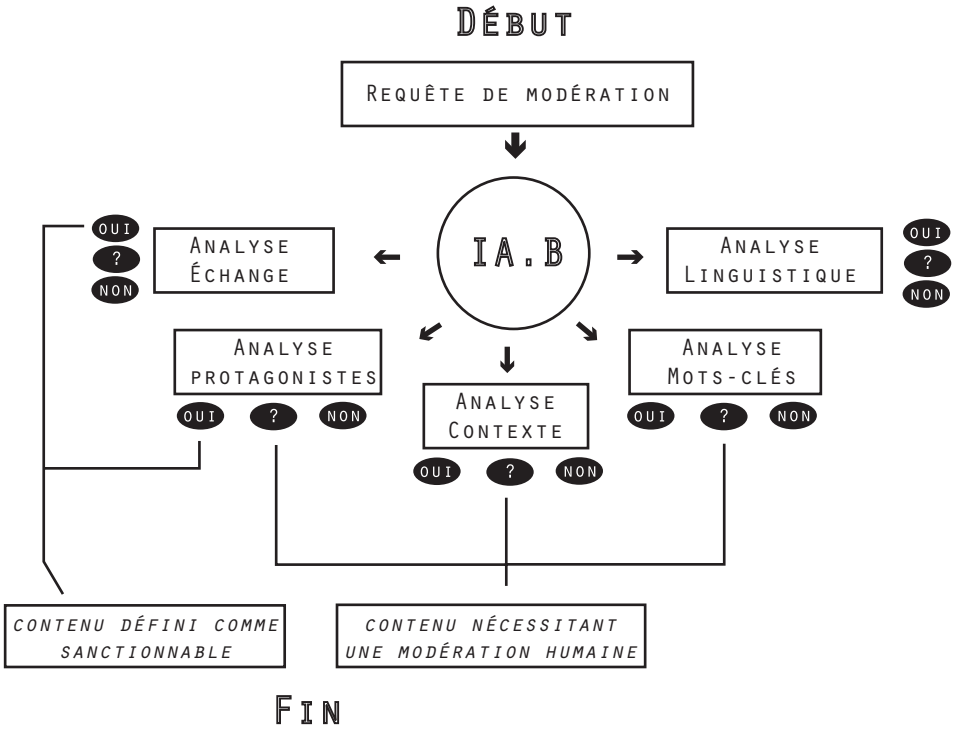
L'analye de l'échange s'intéresse aux protagonistes et à leurs antécédents (temporalité de l'échange, contexte actualité lors de l'échange, nature des acteurs et arrivés dans l'échange etc.) Le but est de déceler des comportements récurrents et de les différencier (trolls // injurieux) mais aussi de comprendre les possibles escalades et les acteurs en causes à des temps donnés.

L'analyse des protagonistes s'intéresse aux humains qui conversent autour du contenu. Cette analyse concerne aussi bien leurs antécédents mais aussi leurs réactions. Ainsi sont pris en compte les like et dislike émit par la communauté d'utilisateurs et les requêtes de modération envoyés.

L'analyse des mots clés consiste en la recherche dans la base de donnée de modération de terme récurrent pouvant porter à confusion. Exemple, les insultes onomatopée type « merde », « putain », « chier ».

3. Chaque analyse rend un résultat : oui (ce contenu est délictueux), non (ce contenu n'est pas délictueux), (ce contenu reste) incertain.
>> Si les incertitudes sont supérieurs à 3 alors l'analyse est envoyé à un modérateur humain.
>> Si l'analyse comporte 2 «non» alors le contenu n'est pas considéré comme sanctionnable.
>> Si l'analyse comporte 2 «oui» alors le contenu est considéré comme sanctionnable.

Oui	Non	Incertain
… < 2 alors sans suite	… < 2 en fonction de *n* oui	… < 3 en fonction de oui et non
… > 2 alors sanction	… > 2 alors sans suite	… > 3 alors envoie à un modérateur humain



4. La machine effectue un rapport global à partir des résultats de chaque analyse.
Ce rapport contient d'une part les résultats mais aussi les tags de chaque analyse. Ces tags permettent de préciser l'analyse et aident à la décision de la sanction. Ainsi si les mot-clés «vulgarité» et «onomatopé» sont présents le résultat final de la machine sera différent d'un contenu avec les mot-clés «vulgarité» et «discrimination».

EXEMPLE DE TAGS

ANALYSE DES PROTAGONISTES

tags associés :

# troll	# engagé
# récidiviste	# sanctionné
# lambda	# vulgaire
# professionnel	

ANALYSE PAR MOT-CLÉS

tags associés :

# vulgarité	# onomatopée
# discrimination	# ponctuation
# thème à risque	# animaux
# memes	# politique

ANALYSE DU CONTEXTE

tags associés :

# ancien	# forum
# récent	# débat
# podcast	# publicité

ANALYSE LINGUISTIQUE

tags associés :

# onomatopée	# ponctuation
# écriture émotive	# poésie
# expression	# proverbe

ANALYSE DE L'ÉCHANGE

tags associés :

# dérision	# confrontation
# assentiment	# conversation
# débat	# avis

5. Une Analyse Centrée est lancée à partir de la synthèse des analyses de la machine A et B.
Le but est de définir la sanction en fonction des paramètres récoltés auparavant.
Par exemple, si la machine a reçu deux oui un non elle étudie la provenance de ceux-ci et prend en compte le résultat et les annotations des analyses. Une *annotation** est un mots-clés associé à un contenu en fonction d'une analyse. Elle permet d'affiner la catégorie du contenu et le type de sanction à lui associer. Les annotations permettent d'alléger la sanction ou au contraire de l'alourdir. Ainsi si le mot-clé « ironie » est associé au contenu la sanction peu être un simple avertissement automatique mais si le mot-clé associé est « xénophobie » alors la sanction peut devenir un envoie automatique du dossier de l'internaute à la jurisprudence et une limitation de son utilisation du web.

6. Les sanctions sont envoyés selon le protocole suivant.
>> Si le contenu est considéré comme fortement sanctionnable(grade 3) + présence d'au moins un mot-clés lourds (relatif au délit ou à la criminalité) => Alors le dossier de l'internaute est envoyé à la juridiction compétente et précisé dans les recherches de profil à risque. Des limitations de l'utilisation de l'espace public Internet sont envoyées à la machine de l'utilisateur et celui-ci est prévenu de la durée de sa mise en quarantaine.
>> Si le contenu est considéré comme fortement sanctionnable (grade 3) mais sans présence de mot-clés lourds (relatif au délit ou à la criminalité) => Alors des limitations de l'utilisation de l'espace public Internet sont envoyées à la machine de l'utilisateur et celui-ci est prévenu de la durée de sa mise en quarantaine.
>> Si le contenu est considéré comme moyennement sanctionnable (grade2) mais avec présence de mot-clés lourds => Alors le dossier de l'internaute est envoyé à la juridiction compétente. L'internaute reçoit un message de mise en cause qui sera utiliser en sa défaveur en cas de récidive.
>> Si le contenu est considéré comme moyennement sanctionnable (grade2) sans aucune présence de mot-clés lourds => Alors l'internaute reçoit un message d'avertissement et de rappel des pratiques sociales sur Internet.
>> Si le contenu est considéré comme faiblement sanctionnable (grade 1) mais avec présences de mot-clés lourds => Alors l'internaute est prévenu de l'ouverture d'un dossier de modération le concernant. Il reçoit des message s d'avertissement et de rappels des pratiques sociales sur Internet.
>> Si le contenu est considéré comme faiblement sanctionnable (grade 2) mais sans présence de mot-clés lourds => Alors l'internaute reçoit des messages d'avertissement et de rappel des pratiques sociales sur Internet.

ÉTUDE DE CAS

Voici un commentaire extrait de la plateforme Youtube. La vidéo dont parle le commentaire est « Pénis : Inclinez-vous ! » de SolangeTeParle. La machine B analyse ce commentaire ainsi :

Analyse par mot-clés

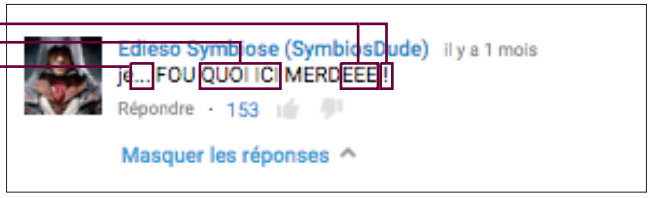
tags associés :
vulgarité

OUI (sanctionable)

Analyse linguistique

tags associés :
écriture émotive
ponctuation

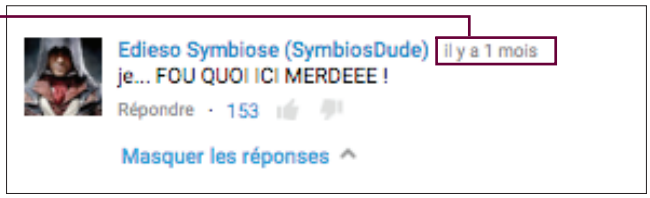
Incertain



Analyse du contexte

tags associés :
podcast
récent

Incertain



Analyse de l'échange

tags associés :
dérision

Non (pas sanctionable)

Analyse des protagonistes

recherche de dossier de modération

tags associés :
lambda

Non (pas sanctionable)

ÉTUDE DE CAS

Voici un commentaire extrait de la plateforme twitter. Le commentaire a été envoyé à la modération. La machine A n'a rien detecté. Elle a cependant transmis la modération à la machine B.

Analyse par mot-clés

tags associés :
animaux
catégorie idéologie

Non (pas sanctionable)

Analyse linguistique

tags associés :
expression /proverbe
négation

Incertain

Analyse du contexte

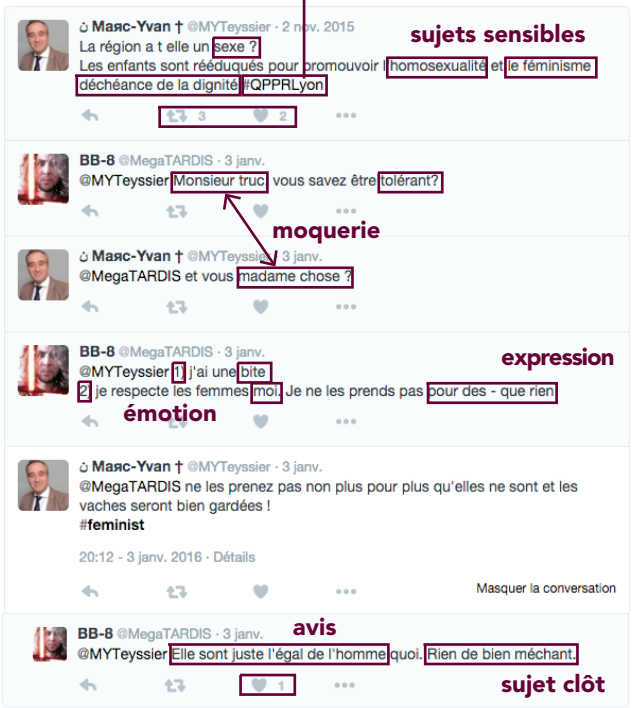
tags associés :
ancien
twitter
conversation

Non (pas sanctionable)

Analyse de l'échange

tags associés :
débat
thème à risque
échange d'idée
politesse

Incertain



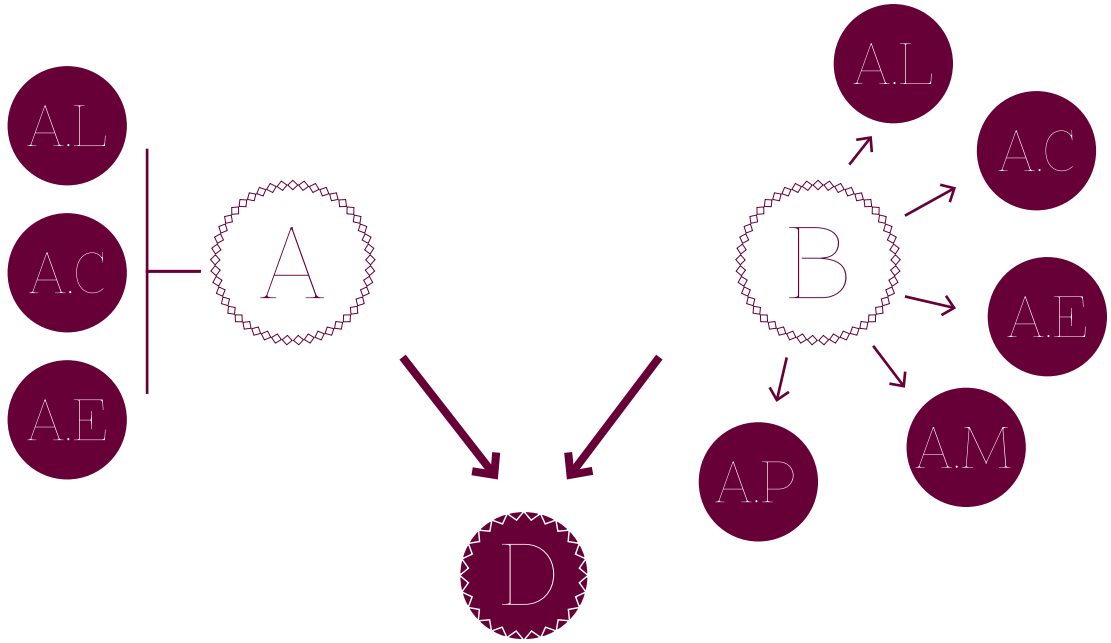
Analyse des protagonistes

Incertain

tags associés :
engagé
vulgaire

Ce cas est assez particulier. La machine ne reconnaîtra pas un comportement abusif à cause de la subtilité de l'échange. Elle transmettra donc la modération à un humain en spécifiant dans son rapport la dimension de débat idéologique. Elle proposera au modérateur d'envoyer au moins un rappel des règles sociale en ligne pour le 2nd utilisateur.

RESSOURCES TECHNIQUES



Sur le plan technique, le système se base sur l'utilisation conjugué du LCP Net* (approche sémantique) et de la logique Floue*. Le but est d'amener le système de modération à devenir un bot informatique* capable de faire de l'analyse de données textuelles. Ainsi l'analyse par mots-clés se rapproche des méthodologies de lexicométrie, et les analyses linguistique et des échanges suivent des méthodes de textométries et de logométrieDans ce contexte il est impératif d'avoir un système structuré, rapide mais aussi très souple.

Le LCP Net permet d'optimiser les analyses complèxes liées à la langue tandis que la logique floue assure une souplesse dans les calculs par l'intégration des incertitudes et l'adatation à de multiples chemins logiques. Par ailleurs, le mode LCP Net grâce à son utilisation du QoS (Qualité de Service) assure de bonnes confitions de traffic des données notamment en terme de débit, de délais de transmission et de variation des latences. Dans la mesure où deux machines fonctionnent en simultanée sur un même objet à modérer, la rapidité de calcul de chacune d'elle est essentiel. Le langage LCP Net intègre aussi des principes de la logique floue dans sa structure ce qui lui permet d'être très adapté à l'analyse des contenus à modérés sur les plans linguistiques mais aussi de nature (image figée, vidéo, texte etc.).

La logique Floue, elle, est primordiale dans les prises de décision de la machine grâce aux nuances qu'elle émet. En effet, elle est présente dans les étapes d'association entre un résultat donnée et les annotations à lui attribuer. Son système plus souple et réactif basé sur *le degré de vérité* plutôt que sur un résultat binaire permet d'affiner les analyses des deux machines. Son imprecision, loin de truquer les résultats des analyses, est un atoût qui fait varier sensiblement les décisions finales et par la même les sanctions. Ainsi la logique floue est particulièrement pertinente dans un système de modération automatique mais qui souhaite immiter les prises de décision humaines.

CONCLUSION

Ainsi donc, le système de modération amélioré imaginé dans cette recherche se base sur une organisation complémentaire de deux protocoles souples fonctionnant principalement sur la logique floue et le LCP Net. Mimant les prises de décision humaine et automatisant la sanction qu'il calcule adéquat, le bot informatique permet d'alléger la charge de travail des modérateurs, de construire un système pédagogique de sanction sur Internet et de replacer l'espace numérique dans un contexte d'intéraction sociale.

Pourquoi une modération à deux machines ?
Les deux intelligences artificielles permettent une analyse plus fine et donc plus sûre des contenus. Sans remplacer la finesse de jugement d'un humain, la modération automatisé sert essentiellement à gérer la masse d'utilisateur par des rappels de savoir-vivre en ligne mais aussi des sanctions prédéfinis et modifiables par les modérateurs humains. Le fait d'avoir deux machines affine le niveau de fiabilité de la sanction. En effet les protocoles de chacune d'entre elles ne sont pas linéaires. Ils suivent soit des analyses parallèles reliées soit des analyses simultanées distinctes. Le fonctionnement de ces analyses multiples est un gage de subtilité dans les résultats obtenus. Et cela est essentiel dans un système automatique qui agit avec conséquence dans le quotidien d'utilisateur humain.

Pourquoi ces sanctions en grade ?
Le système de sanction imaginé est basé sur le principe de mort sociale plutôt que sur la seule condamnation juridique. En effet, en sanctionnant fortement les utilisateurs qui postent des contenus abusifs par une mise en quarantaine et une coupure ou une limitation de l'accès à un Internet, le système devient un régulateur social. Mais il ne peut pas être qu'un simple garde fou. C'est pourquoi le système de grade permet d'introduire de la pédagogie dans la modération en avertissant un utilisateur de certaine de ses dérives - notamment sur le plan de la légereté de langage. Le système mime les comportements sociaux et rappelle que le web est un espace public soumit au droit et à la loi mais aussi aux convenances sociales. Aussi l'atoût de ce système est de rendre visible les comportements et de les signaler aux autres utilisateurs. C'est aussi pour cette raison qu'il y'a une distinction entre les sanctions annotées par des mots-clés à caractère délictueux ou criminel et les sanctions annotées de mots-clés annodins.

Est-ce qu'un système uniquement sécuritaire fonctionnerait ? Et l'éthique dans tout ça ?
Évidemment ce présent document est une réponse à court terme pour des questions sociales bien plus profondes. La mouvance sécuritaire qu'a pris nos sociétés, si elle s'argumente de nombreuses belles et clairvoyantes justifications n'en demeure pas moins criticables sur le plan éthique. En ce sens les débats menés par la Quadrature du Net sont des plus légitimes. Aujourd'hui le World Wild Web n'a que vingt-cinq ans et est encore en grand chantier. Les enjeux de sécurité des données sont vite passés dans les réponses pratiques avant la protection des utilisateurs et cette lacune se vit très fortement alors que la Neutralité du Web est directement attaqué par tout un tas de lois européenne et française. Ainsi donc si un tel système pourrait être fonctionnel, il n'en reste pas moins éthiquement criticable et même dans la pratique assez limité.

Comment pourrait on faire de la modération autrement ?
Un autre réponse possible à l'amélioration de la modération est toute simple dans la mesure où elle s'affranchit d'outil de modération. Aussi saugrenu que puisse sembler être cette réponse elle est *viable, fiable et acceptable* et repose, comme ce précédent axiome, sur des logiques de design. Aujourd'hui les principaux comportements abusifs sur Internet sont induit par les interfaces des utilisateurs. Celles-ci étant elles-mêmes basées sur des outils informatiques, il suffirait de changer ces derniers pour transformer en profondeur les habitus des internautes. Par exemple des interfaces qui, à l'inverse de Twitter, empêchent les écrits numériques de faire moins de 1500 caractères, résoudraient bon nombres de commentaires émotifs , impulsifs et irréfléchis qui pullulent sur les média sociaux.

De même, des comportements qui correspondent plus à des fonctionnalités orales n'ont pas besoin d'être archivé *ad vitam eternam*. Si les textes de moins d'une ligne étaient supprimées au bout de x temps, alors beaucoup des commentaires abusifs n'agresseraient pas la vue des utilisateurs. Ce dernier point repose sur un des grands enjeux d'Internet : la tracéologie. À l'instar du plurilinguisme, de la géographisation du web, de la protection des données et des utilisateurs, la tracéologie est un objet de recherche assez important mais très laborieusement appliqué. Elle s'intéresse à l'étude des traces numériques en tant qu'objet patrimonial (ce qui induit la création d'une Histoire d'Internet) mais aussi en tant qu'élement dynamique de mise en archive. Or un des principaux élements de la mémoire est sa capacité à oublier. Il n'y a, en effet, aucun processus de mémorisation sans sélection. Et c'est là que réside une des limites du web actuel car celui ci ne connaît que le stockage. L'archivage n'étant quasiment pas organisé dans une optique de construction du web mais dans une logique d'accumulation, il entraîne une sur-utilisation des données (avec la fascination pour les BigData), de l'énergie nécessaire à les conserver et du temps d'attention des utilisateurs pour réorganiser leurs propres traces numériques.

Ce présent document n'avait pas pour ambition de solutionner les nombreux chemins que le web, en tant que nouvel espace d'intéractions sociales, offre mais d'essayer de résoudre concrêtement par la logique le problème contemporain qu'est la modération en ligne. Les nombreuses limites techniques ne me permettent pas, à l'heure actuelle, de définir de la pertinence de cette proposition ni de sa viabilité opérationnelle. En tant que pure réflexion conceptuelle sans calcul détaillée ni test appliquée, cette recherche est à prendre pour ce qu'elle est, un processus ouvert qui n'attend qu'à être mis en oeuvre.

En vous remerciant pour votre lecture…