



“SafeTweet”

CAHIER DES CHARGES

AUGER Nathan
DURAND Pierre
LOPEZ Julio
NOUVELIÈRE Benjamin

Table des Matières

| | |
|------------------------------------------------|----------|
| 1- Contexte | 2 |
| 2- Objectif | 2 |
| 3- Fonctionnalités | 3 |
| 4- Maquettes et scénarios d'utilisation | 4 |

1. Contexte

L'accès aux informations sur internet devient facile et instantané grâce aux moyens techniques dont nous disposons (internet illimité, 4G) et à la démocratisation de l'usage des objets connectés et des réseaux sociaux. Cependant, un phénomène mondialement connu est en train de bouleverser notre comportement : les fake news. L'impact de fake news a été étudié sur plusieurs aspects de notre vie quotidienne sur les réseaux sociaux : l'apprentissage, le raisonnement, la prise de décision, etc.

En tant qu'étudiants en Master première année d'informatique à Le Mans Université et dans un cadre R&D, notre travail consiste à étudier une nouvelle approche permettant de détecter les fake news.

Les pistes à creuser sont multiples : l'intelligence artificielle, le moteur d'inférence, l'analyse sémantique des données, etc.

Et si nous ne nous arrêtons pas qu'aux fake news ? Hoax, Internet Scum, Cyber Security,... ils nous concerne aussi.

2. Objectif

L'objectif sera de recréer une application Twitter pour mobile dont la fonctionnalité majeure sera de pouvoir détecter les tweets à caractère diffamatoire ou harcelant.

3. Fonctionnalités

- Lors de la première ouverture de l'application, l'application demande à l'utilisateur de se connecter à son compte twitter.
- L'utilisateur sera en mesure de consulter sa timeline twitter et de se déconnecter (cette action entraînant le retour à la page de connexion).
- Un système d'analyse des tweets sera mis en place, celui-ci permettra de détecter la visée du tweet. Il s'agit là de voir si le tweet contient de la diffamation ou du harcèlement.
- En utilisant un code couleur et des logos appropriés, un système de notation simple et intuitif permettra à l'utilisateur d'être informé de la visée du tweet.
- L'utilisateur pourra enrichir le système en signalant un tweet, dans le cas où le système n'a pas détecté de harcèlement ou de diffamation. Dans le cas d'un tweet étant marqué (d'un harcèlement ou d'une diffamation), l'utilisateur pourra démentir le jugement du système.
- L'utilisateur pourra signaler l'auteur d'un tweet.
- L'utilisateur aura la possibilité de filtrer les tweets qui s'affichent dans sa timeline:
 - Afficher tous les tweets
 - Cacher ceux comportant du harcèlement
 - Cacher ceux comportant de la diffamation
 - N'afficher aucuns des tweets contenant du harcèlement et de la diffamation

4. Maquettes et scénarios d'utilisation




| <i>Pictogramme</i> | <i>Légende</i> |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
|  | <i>Signifie qu'un tweet ne présente aucune offense à l'égard de l'utilisateur.</i> |
|  | <i>Signifie qu'un tweet présente une offense à l'égard de l'utilisateur, qu'il soit vrai (harcèlement) ou faux (diffamation).</i> |
|  | <i>Signifie qu'un tweet présente une forte offense à l'égard de l'utilisateur, qu'il soit vrai (harcèlement) ou faux (diffamation).</i> |



Fig 0. Pictogrammes représentant le niveau d'offense des tweets

Ces pictogrammes assurent une information direct à l'utilisateur sur l'offense du message, placé avant le texte, ils laissent la possibilité à l'utilisateur de lire ou non le tweet.



Fig 1. Page de connexion de l'application

C'est ici la page lors de la première ouverture de l'application. L'utilisateur est alors invité à se connecter à son compte Twitter. La connexion est conservée d'une session à l'autre. De ce fait, la prochaine fois que l'utilisateur ouvrira l'application il ne passera pas cette page mais sera directement redirigé vers la page principale.

Si l'utilisateur décide de lui-même de se déconnecter alors cette page s'affichera de nouveau pour qu'un nouvel utilisateur puisse se connecter.

Un lien cliquable "cliquer ici pour la notice" enverra l'utilisateur vers une nouvelle page présentée (**Fig 0**) précédemment, afin d'informer l'utilisateur sur l'application.

Un deuxième lien cliquable “Mot de passe oublié” enverra l'utilisateur sur une page web https://twitter.com/account/begin_password_reset, celle-la meme où l'utilisateur peut demander la réinitialisation de son mot de passe.



Fig 2. Page principale de l'application, l'utilisateur consulte sa timeline.

La page principale de l'application. C'est sur cette page que l'application s'ouvrira lorsque l'utilisateur aura renseigné ses identifiants de connexion (la connexion au compte étant conservée d'une session à l'autre).

Ici l'utilisateur consulte sa timeline en faisant défiler les tweets tout comme il le ferait sur une application Twitter classique.

En haut à droite de chaque tweet, l'utilisateur peut consulter la “dangerosité” du tweet via un symbol et un code couleur.

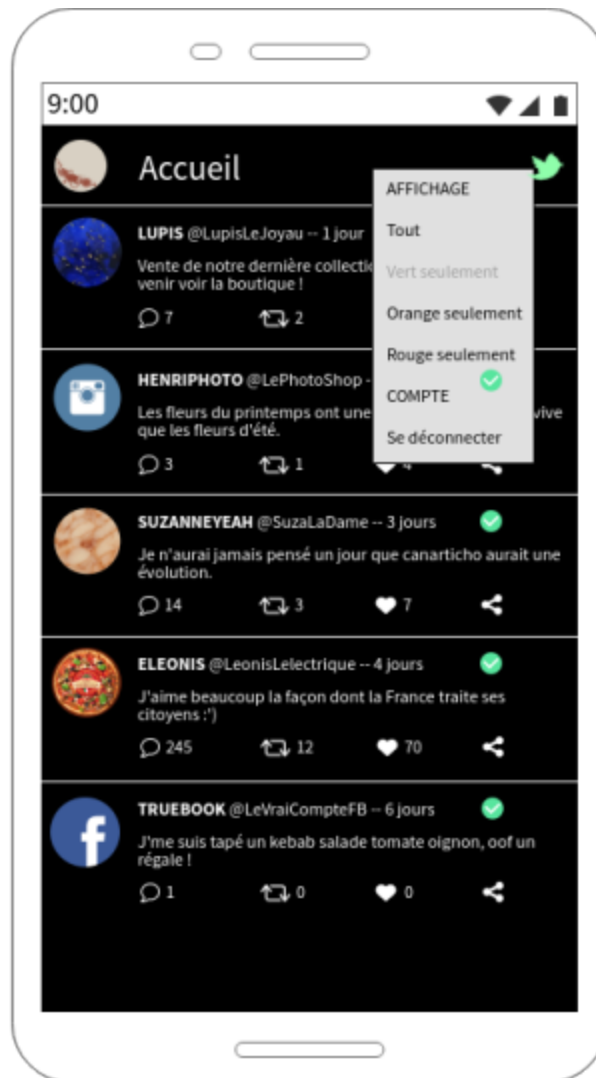


Fig 3. Paramètres de l'application, l'utilisateur peut choisir de masquer certains tweets. Il peut aussi se déconnecter.

Les paramètres de l'application, l'utilisateur peut se déconnecter, il sera alors redirigé vers la page de connexion (l'information de connexion entre les sessions est alors supprimée).

L'utilisateur a, de plus, la possibilité de filtrer l'affichage des tweets : masquer les discriminations et/ou le harcèlement.



Fig 4. Signalement d'un tweet

En cliquant sur le logo d'évaluation du tweet, l'utilisateur peut alors connaître la raison du signalement du tweet (harcèlement, diffamation ou safe).

Le bouton "Démentir" permet de signaler au système qu'il s'est trompé sur l'évaluation du tweet.

Le bouton "Signaler" permet de signaler l'auteur du tweet auprès de Twitter.

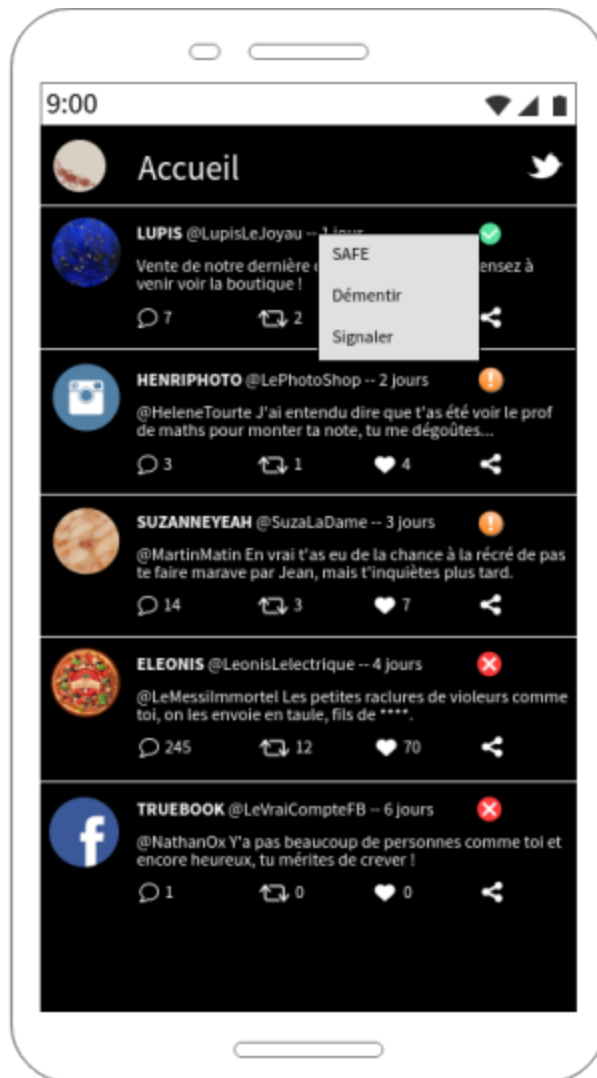


Fig 5. Même exemple que précédemment mais dans le cas d'un tweet classé comme "safe".

Ici, l'utilisateur peut avertir le système que le tweet jugé comme "safe" contient potentiellement une forme de harcèlement ou de discrimination. L'auteur peut aussi être signalé comme dans le cas précédent.