

# Projet VBA/RExcel: Analyse du sentiment - Notice

*F. Boizard, A. El Khaloui, K. Lescoet, M. Le Tertre*

*Lundi 17 Novembre 2014*

## Contents

<b>1</b>	<b>Contexte et Problématique</b>	<b>1</b>
<b>2</b>	<b>Présentation de l'application et de son fonctionnement</b>	<b>2</b>
2.1	Utilisateurs visés . . . . .	2
2.2	Configuration requise . . . . .	2
2.3	Principes de fonctionnement . . . . .	3
<b>3</b>	<b>Guide pas à pas</b>	<b>3</b>
3.1	Profil émotionnel avec un mot clé . . . . .	3
3.2	Comparaison de profils émotionnels . . . . .	6
3.3	Suivi temporel du profil émotionnel . . . . .	6
3.4	Autres fonctionnalités . . . . .	7

## 1 Contexte et Problématique

La perception qu'ont les consommateurs d'une marque est une grandeur difficile voir impossible à mesurer directement. Elle n'en demeure pas moins essentielle, et représente une donnée extrêmement intéressante pour les entreprises qui se soucient de l'image de leurs marques.

Chaque marque en effet génère, via les différents stimuli qu'elle affiche des sentiments et des émotions chez les consommateurs. L'analyse de ces derniers peut permettre à la marque tout d'abord de juger de sa popularité, mais aussi de jauger l'évolution de son image de marque, éventuellement dans l'optique de l'améliorer ou de l'adapter à ses objectifs.

Twitter est un réseau social particulier, en cela qu'il est principalement utilisé pour donner son avis de manière rapide et concise sur les différentes expériences que l'on vit dans sa journée. Plus particulièrement, il est très utilisé pour rendre compte de l'expérience que l'on a avec des marques ou des produits ; Et c'est précisément ce qui nous intéresse.

Nous nous proposons de faire l'analyse textuelle des tweets se rapportant à une marque, dans l'optique d'en dresser le profil émotionnel. Nous pourrions également ajouter des fonctions de comparaisons de profils émotionnels.

Se servir des tweets rédigés en français reviendrait à se limiter à une population minime (urbaine, masculine, de CSP élevée et typiquement "high tech"), du fait du faible taux de pénétration de Twitter en France, qui est vu comme un réseau social compliqué, plutôt sérieux et principalement à orientation professionnelle. Nous faisons le choix de ne traiter que les tweets écrits en anglais.

En effet, Twitter est d'utilisation bien plus commune dans les pays anglo-saxons, mieux répartie en termes de catégories socio-professionnelles, d'âge et de sexe. Cela permet de plus de ne pas poser de contraintes géographiques de provenance des tweets. On travaille ainsi de manière plus générale, sur des données issues de plusieurs pays. Le revers est que les marques sur lesquelles on travaillera seront des marques internationales.

## 2 Présentation de l'application et de son fonctionnement

### 2.1 Utilisateurs visés

L'application intéressera sans doute les professionnels du marketing désireux de suivre la perception de leur marque, ou de suivre l'impact d'une campagne publicitaire sur celle-ci par exemple. Elle a également un intérêt pour les personnes désirant acheter des parts de capital social dans une entreprise, en tant que moyen de vérifier que les marques de la firme sont suffisamment populaires et bien perçues pour lui garantir un revenu futur. En effet, on se doute qu'une entreprise associée principalement à des émotions négatives risque fort de voir ses ventes, et a fortiori ses revenus diminuer. L'utilisation de cette application ne demandera aucune compétence statistique spécifique.

L'application n'est cependant pas réservée à l'étude marketing. Elle peut être utilisée pour étudier le profil émotionnel de n'importe quel autre mot-clé. On peut ainsi s'intéresser à des sujets comme la politique, l'actualité, l'économie, etc.

### 2.2 Configuration requise

#### 2.2.1 Matériel

De par son fonctionnement, l'application requiert tout d'abord une connexion internet. Ensuite, vu le nombre de packages qu'elle charge et le volume de données qu'elle traite, il faut une machine puissante en termes de mémoire vive (Un minimum de 4Go de RAM permet un fonctionnement optimal).

#### 2.2.2 Logiciels

L'application requiert les programmes suivants:

- **Pack Office** (dont **Excel**), versions 2007 et supérieures avec Macros activées.
- Une installation fonctionnelle de **RExcel** et **Statconn**.
- Une version récente de **R** (Version 3.1.2 et supérieures)

L'installation des packages **R** requis est directement implémentée dans l'application. Voici une liste des packages requis:

- **sentiment** : permet de faire l'analyse textuelle du corpus de texte, en associant un sentiment aux mots reconnus comme faisant partie des dictionnaires du package.
- **twitterR** : permet d'accéder à l'API Twitter et de l'interroger.
- **plyr** : est utilisé lors de la phase de "nettoyage" du texte. Il facilite le processus en le divisant en tâches simples.
- **wordcloud** : permet de tracer la sortie nuage de mots.
- **ggplot2** : package qui nous servira de moteur graphique, choisi à cause de son esthétique. Sera potentiellement abandonné au profit du device **R** de base.
- **RColorBrewer** : librairie permettant une gestion étendue des couleurs, comme la création de palettes de dégradés. Ce package n'est pas indispensable non plus.
- **devtools**: requis pour l'installation du package **sentiment**.

L'installation des packages est totalement transparente et se fait, si ce n'est déjà fait, au lancement du fichier xlsx contenant la macro.

*Remarque: chacun de ces packages en appelle éventuellement d'autres, dits dépendances.*

## 2.3 Principes de fonctionnement

Le principe de fonctionnement de l'application suit les étapes suivantes:

- Téléchargement de données depuis les serveurs de <http://Twitter.com>.
- Mise en forme des données ("nettoyage").
- Analyse textuelle des données, attribution d'une polarité (positive, negative ou neutral) et d'une émotion (anger, disgust, joy, fear, sadness, surprise) par classification naïve Bayésienne à chacun des tweets du corpus importé.
- Mise en forme des données sous le format `data.frame`, construction de tableaux de contingence.
- Représentation graphique des données selon le type d'analyse désiré. L'application propose d'exporter les sorties graphiques dans un document PDF.

Sachant que les données sont foncièrement différentes d'un type d'analyse à l'autre, l'application propose trois boutons d'importation, un pour chaque analyse.

Lorsqu'on lance le fichier xlsx, le menu qui permet de lancer des commandes et d'exécuter des analyses apparaît en tant qu'onglet dans le ruban Excel

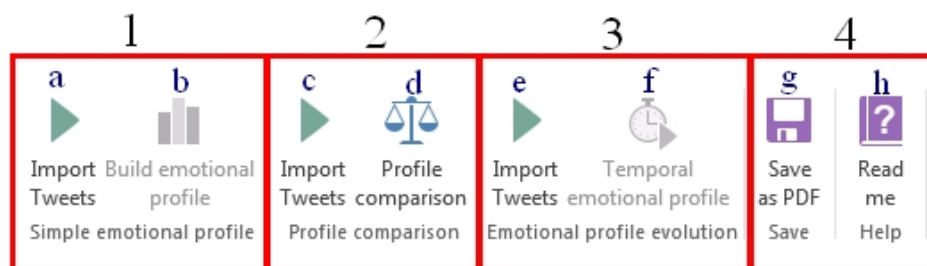


Voici à quoi ressemble le ruban:



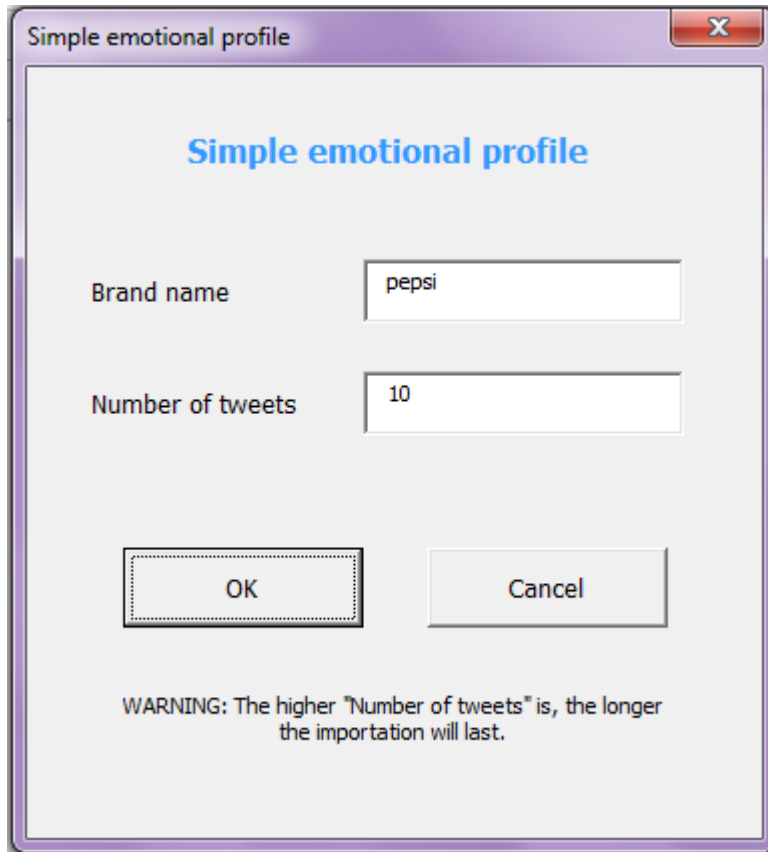
## 3 Guide pas à pas

### 3.1 Profil émotionnel avec un mot clé



La première fonctionnalité proposée par l'application est celle qui permet, à partir d'un mot-clé (i.e un nom de marque) et un nombre  $n$  de tweets de dresser le profil émotionnel récent. Cette procédure se base sur les  $n$  derniers tweets à propos du mot-clé entré. Il s'agit donc de dresser un bilan émotionnel à l'instant  $t$  (1).

- La première étape consiste à importer les tweets à propos d'une marque (a).



Simple emotional profile

Simple emotional profile

Brand name

Number of tweets

WARNING: The higher "Number of tweets" is, the longer the importation will last.



- Les paramètres à fournir sont les suivants:
  - Un mot-clé: la marque d'intérêt.
  - Un nombre de tweets. Le temps que requiert l'importation est fonction de ce paramètre.
  - Si oui ou non les tweets importés doivent impérativement avoir été compris en termes d'émotion par l'application. Les tweets non "compris" sont alors supprimés.
  - Si oui ou non les retweets doivent être pris en compte et importés dans le corpus.

\*\* Attention, lorsque l'on importe uniquement les tweets qui ont été "compris" par la macro, en supprimant les retweets, la taille du corpus est drastiquement réduite. Il faut donc importer des tweets en plus grand nombre pour avoir suffisamment de matière à traiter.\*\*

\*\* Il faut noter que la suppression des retweets ne doit pas être systématique, car elle est directement liée à la notion de "buzz". Un tweet qui a été retweeté plusieurs fois est un tweet qui a été jugé comme "bon", "exact" et "juste" par les utilisateurs.\*\*

- L'application affiche les tweets importés afin que l'utilisateur puisse s'assurer qu'ils ne comportent pas d'anomalie et que l'importation s'est correctement déroulée. A noter qu'il arrive que des caractères spéciaux se glissent dans les tweets importés, malgré le nettoyage que nous leur feront subir; cela est à mettre en lien avec l'utilisation croissante de "smileys", notamment dans le cas des utilisateurs mobiles.
- On demande ensuite dans un second temps de représenter le profil émotionnel **(b)**. Les sorties sont les suivantes:
  - Le profil émotionnel.
  - Le profil de polarité.
  - Un nuage de mots avec les mots les plus utilisés dans le corpus de tweets à propos du mot-clé, classés et colorés par émotion.

### 3.2 Comparaison de profils émotionnels

Cette seconde fonctionnalité permet de faire les profils émotionnels pour 2 à 5 mot-clés, et de représenter ces derniers de manière conjointe (2).

Emotional Profile Comparison

**Emotional profile comparison**

Number of brands: ☐ 2 ☐ 3 ☒ 4 ☐ 5

Brand N°1:

Brand N°2:

Brand N°3:

Brand N°4:

Brand N°5:

Number of tweets:

OK Cancel

- L'étape première consiste assez naturellement à faire les importations de tweets, pour les différentes mots-clés d'intérêt. Il est aussi possible de ne pas importer les tweets non "compris" (c). Le nombre de tweets entré en paramètre correspond au nombre de tweets importés par marque.
- Les tweets importés sont, comme précédemment affichés pour vérification.
- Cliquer sur le bouton "comparaison de profil" construit et affiche les sorties graphiques comparatives pour les émotions et la polarité (d). Les sorties sont des graphiques comparatifs des profils émotionnel et de polarité
- L'import en PDF est possible (g).

### 3.3 Suivi temporel du profil émotionnel

Cette dernière fonctionnalité, permet d'intégrer un aspect évolutif au profil émotionnel. L'argument n est alors le nombre de tweets par jour utilisés. Elle requiert l'ajout de deux dates( début, et fin) au format AAAA-MM-JJ. Cette recherche ne peut se faire que sur les 9 derniers jours, du fait des données fournies

par Twitter via son API. Les sorties de cette fonctions permettent de visualiser l'évolution des émotions et de la polarité des tweets écrits à propos d'une marque. Ceci peut se révéler très utile pour évaluer l'impact d'une campagne de publicité au cours du temps. Les étapes à suivre sont similaires à ce que l'on fait pour les fonctionnalités précédemment décrites (3).

**Temporal emotional profile**

**Brand name**

**From :**  **To :**

**Number of tweets**

**Calendar (November 2014):**

novembre 2014						
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7

**Today: 20/11/2014**

- Comme précédemment, on procède à une importation (e) puis à une représentation des données (f), qui sont ici sous la forme de séries temporelles.
- Lors de l'importation, le nombre de tweets à importer correspond à un nombre par jour.
- Les sorties consistent en deux graphiques d'évolution des profils émotionnel et de polarité entre les deux dates choisies.

### 3.4 Autres fonctionnalités

- Il est possible d'exporter les sorties graphiques de la macro dans un document pdf (g).
- Le code R derrière la macro Excel a été conçu de manière à éviter les arrêts d'exécution dus à des incompatibilités d'encodage ou à des caractères spéciaux non supportés par R. Un travail conséquent a été fait dans cette direction, afin que l'application soit opérationnelle en permanence. Pour ce faire, la connexion par défaut à l'API Twitter a été contournée pour effectuer une connexion directe, moins sensible aux petites erreurs et irrégularités.

- Par nature, le programme R derrière la macro est lent à exécuter dès que l'on travaille sur des corpus un tant soit peu volumineux. Interroger les bases de données Twitter prend du temps et traiter le texte requiert de la puissance de calcul. Nous avons mis un point d'honneur à optimiser l'exécution du script, en réduisant le nombre d'opération gourmandes en ressources au minimum, et en limitant les appels de fonctions. L'application privilégie les opérations en local et a été construite de manière parcimonieuse.
- Il est **très important** de faire attention au ton utilisé dans les tweets. En effet, l'application ne comprend pas l'ironie ni l'humour. Il arrive en effet que des tweets humoristiques circulent à propos d'une marque, et c'est précisément la raison pour laquelle un moyen de contrôler les tweets avant de les analyser a été implémenté. Il est donc toujours bon de **contrôler rapidement, par une lecture rapide** que les tweets à propos d'une marque sont pertinents.



[ LMA<sup>2</sup> ]

# Sentiment profiling Toolbox



*Spe STAT 2014/2015*