


Surveillance en temps réel de la e-réputation des entreprises

RAPPORT DU PROJET DE BIG DATA



GBONGBON Roméo
KOUAM MATCHIM LORRAINE Stella
TCHONGOUANG DJOMO Gatien Junior
ENOBIL FRANKLIN Claire

INTRODUCTION

À l'ère du numérique, la présence en ligne des entreprises est scrutée en continu par une audience globale. Les médias sociaux, en particulier, sont devenus un acteur majeur dans la réputation des marques, impactant directement leur succès et leur pérennité. Dans ce contexte dynamique et souvent impitoyable, la capacité à évaluer et à répondre de manière précise aux besoins des consommateurs est devenue une nécessité opérationnelle. Raison pour laquelle nous avons choisi d'implémenter un système d'analyse de sentiments en temps réel dédié à la surveillance et à la gestion de la e-réputation des entreprises sur les médias sociaux.

L'objectif de cette solution est multiple. Premièrement, il s'agit de détecter de manière précoce les problèmes de réputation qui peuvent émerger afin de permettre une intervention rapide et efficace ; minimisant ainsi les dommages potentiels.

Deuxièmement, ce système est conçu pour repérer les occasions susceptibles d'améliorer l'image de la marque et d'établir un dialogue constructif avec les clients.

Troisièmement, dans le cadre de la gestion de crise, la réactivité est essentielle : une entreprise doit pouvoir répondre avec agilité aux critiques et aux incidents pour maintenir sa réputation.

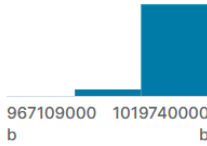
Enfin, l'analyse des tendances grâce à la surveillance en temps réel des médias sociaux permettra d'anticiper les changements dans l'opinion publique et d'adapter les stratégies de communication pour rester en phase avec les attentes des consommateurs.


La réalisation de ce système est un pas en avant stratégique car elle assure que la réputation numérique de l'entreprise soit non seulement protégée mais également développée, afin de contribuer à la réalisation de ses ambitions à long terme.

PRESENTATION DU PROJET

Extraction et Prétraitement des Données

Dans le cadre de ce projet, nous avons extrait un ensemble de données disponible sur Kaggle, une plateforme réputée pour ses vastes ressources en matière de données. Le set de données, comprenait plusieurs colonnes :

# id	text	timestamp	source	symbols
 967109000 1019740000 b b	25841 unique values	26948 unique values	bibeypost_stock 3% whatsonthorold2 3% Other (26487) 93%	455 unique values
1.01971E+18	VIDEO: "I was in my office. I was minding my own business..." -David Solomon tells \$GS interns how h...	Wed Jul 18 21:33:26 +0000 2018	GoldmanSachs	GS
1.01971E+18	The price of lumber \$LB_F is down 22% since hitting its YTD highs. The Macy's \$M turnaround is still...	Wed Jul 18 22:22:47 +0000 2018	StockTwits	M
1.01971E+18	Who says the American Dream is dead? https://t.co/CRgx19x7sA	Wed Jul 18 22:32:01 +0000 2018	TheStreet	AIG

Detail Compact Column					8 of 8 columns	
source		▲ symbols	▲ company_names	🔗 url	✓ verified	
ibeypost_stock	3%	455 unique values	462 unique values	[null] 22%	 <div> true 363 1% false 28.1k 99% [null] 4 0% </div>	
rhatsonthorold2	3%			http://binance.com/... 3%		
Other (26487)	93%			Other (21199) 75%		
oldmanSachs		GS	The Goldman Sachs	https://twitter.com/i/web/status/1019696670777503745	TRUE	
tockTwits		M	Macy's	https://twitter.com/i/web/status/1019709091038547968	TRUE	
heStreet		AIG	American	https://buff.ly/2L3kmc4	TRUE	

Donc nous avons extrait les informations les plus pertinentes pour notre analyse : l'identifiant du commentaire (**id**), le texte du commentaire (**text**), le symbole boursier associé (**symbols**) et le nom de l'entreprise concernée (**company_names**).

Pour nettoyer et préparer les données, nous avons utilisé Talend, un outil d'intégration de données robuste et flexible. Nous avons aussi mis en œuvre un processus d'auto-incrémentation pour les identifiants afin de garantir l'unicité de chaque enregistrement. Nous avons ainsi obtenu en sortie un fichier **csv** que l'on a nommé **out20.csv**.

Cette étape de prétraitement était essentielle pour assurer l'intégrité des données avant l'analyse.

Script d'Analyse de Sentiments

Le cœur de notre analyse réside dans un script Python (fichier **analyzer.py** dans le projet) personnalisé qui prend en entrée le fichier CSV prétraité : **out20.csv**. En utilisant la bibliothèque Natural Language Toolkit (**NLTK**), le script analyse le sentiment intrinsèque de chaque commentaire. Il attribue ensuite à chaque commentaire une catégorie de sentiment : **positif**, **négatif** ou **neutre**. Et on obtient en sortie un nouveau fichier CSV identique au fichier d'entrée, mais ayant en plus une colonne représentant la catégorie de sentiment ; nous l'avons nommé **out20_with_sentiments.csv**.

Cette classification est basée sur une évaluation algorithmique des mots et des expressions utilisés dans le texte, en tenant compte de leur contexte et de leur connotation.

Visualisation des Résultats

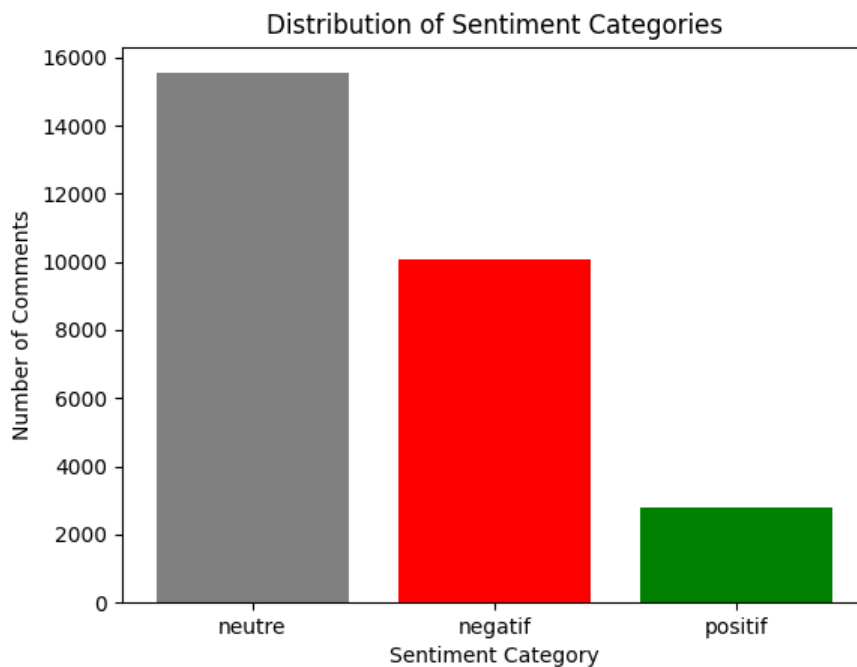
Pour représenter visuellement les résultats de notre analyse de sentiments, nous avons écrit un autre script Python (**dashboard.py**) en s'appuyant sur la bibliothèque Matplotlib. Ce Script prend en entrée notre set de données final **out20_with_sentiments.csv**.

Notre script génère ainsi un graphique à barres qui montre la distribution des sentiments à travers l'ensemble des commentaires. Ce graphique offre une vue d'ensemble immédiate de la tonalité générale des discussions liées aux entreprises sur les réseaux sociaux, mettant en lumière les tendances prédominantes et offrant un aperçu significatif du paysage de la réputation en ligne.

Résultats

L'analyse a révélé que la majorité des commentaires étaient neutres, ce qui peut indiquer une réception modérée ou une absence d'opinions fortes par rapport aux entreprises mentionnées.

Les commentaires négatifs étaient présents en quantité significative, suggérant des domaines potentiels d'amélioration pour les entreprises concernées. Les commentaires positifs, bien que les moins nombreux, reflètent les aspects favorables associés aux marques.



Pour faire fonctionner notre solution, nous devons procéder comme suite :

- Ouvrir le dossier contenant le projet
- Ouvrir l'invite de commande à partir de ce dernier
- Taper la commande : **docker-compose up --build -d**
- Cliquer sur l'image graph.png afin que le graphe s'affiche.

Interprétation

La prédominance de commentaires neutres peut être interprétée de plusieurs manières. Une possibilité est que les utilisateurs des médias sociaux partagent des informations sans exprimer d'émotion forte ou que les algorithmes de classification peinent à discerner des sentiments subtils. Les commentaires négatifs, quant à eux, doivent être scrutés de près car ils pourraient indiquer des problèmes réels ou des malentendus concernant les entreprises. Les commentaires positifs, bien qu'en moindre quantité, sont cruciaux pour construire et maintenir une image de marque positive.

Limites

Les limitations de cette étude incluent la dépendance envers l'exactitude des algorithmes de classification des sentiments et la représentativité des commentaires analysés.

De plus, les sentiments exprimés sur les médias sociaux peuvent être influencés par des événements temporaires ou des campagnes promotionnelles, ce qui peut biaiser les perceptions.

CONCLUSION

En résumé, notre processus d'analyse de données a combiné des techniques de prétraitement de données avancées avec des analyses de sentiments sophistiquées, aboutissant à une représentation graphique claire et précise des sentiments du public. Ce travail illustre la synergie entre des outils de pointe en matière de traitement de données et d'analyse de texte, et démontre notre capacité à extraire des informations précieuses à partir de grands volumes de données non structurées.

Cependant, au cours de l'implémentation de ce projet ambitieux, nous avons été confrontés à divers défis, dont le plus notable a été la conception de l'API destinée à effectuer le streaming et à interagir avec la base de données. Cette étape cruciale, bien que complexe, a été essentielle pour garantir la collecte en temps réel des données provenant des médias sociaux. Malgré ces obstacles, notre équipe a démontré sa résilience et son aptitude à surmonter des problèmes techniques complexes.

Dans l'ensemble, cette initiative a non seulement renforcé nos compétences en matière d'analyse de données, mais a également souligné l'importance de l'adaptabilité et de la résolution de problèmes dans des environnements techniques en constante évolution.

Lien vers le projet : [Claire080/Projet-de-Big-Data: Surveillance en temps réel de la e-réputation des entreprises. \(github.com\)](https://github.com/Claire080/Projet-de-Big-Data: Surveillance en temps réel de la e-réputation des entreprises)