Evaluation Certifiante

Rapport Analytique : Nettoyage, Structuration et Exploration des Tweets SAV Free

Bloc optionnel 2 - CC2.1

Réalisé par :

SGHIOURI Mohammed
SOBGUI Ivan Joel
BOTI Armel Cyrille
BEN LOL Oumar
DIVENGI KIBANGUDI BUNKEMBO Nagui
ELOUMOU MBOUDOU Pascal Aurele

Date : 19 Novembre 2025

Tables de matières

Application en ligne : https://dallosh-analysis.agglomy.com/
Dépôt GitHub : https://github.com/sobgui/dallosh_analysis

Liste des tableaux

Liste des figures

1. Nettoyage et structuration des données

1.1. Suppression des tweets inutiles

Le dataset free_tweet_export.csv contient environ 5 000 tweets bruts. Un filtrage rigoureux isole les tweets pertinents pour l'analyse SAV.

1.1.1. Élimination des retweets

Les retweets automatiques représentent une part significative du dataset. Ces messages ne constituent pas des interactions directes avec Free et sont éliminés via la colonne retweeted_status. Cette suppression permet de se concentrer sur les tweets originaux contenant des demandes, plaintes ou commentaires directs.

1.1.2. Détection et suppression des doublons

La déduplication s'effectue selon plusieurs critères : - **Identifiant unique** : Les tweets avec le même id sont des doublons exacts - **Similarité textuelle** : Distance de Levenshtein pour identifier les tweets quasi-identiques (variations mineures d'espaces, ponctuation) - **Méthode statistique IQR** : Interquartile Range pour détecter et éliminer les outliers textuels (spams, messages aberrants)

Cette approche garantit un nettoyage efficace tout en préservant les tweets légitimes.

1.1.3. Filtrage des tweets hors sujet

Les tweets sont classés comme "hors sujet" s'ils correspondent aux critères suivants : - Messages promotionnels ou publicitaires sans demande client - Réponses internes entre comptes Free (identifiables via les métadonnées in_reply_to) - Tweets humoristiques ou parodiques sans lien avec un problème réel - Messages de spam ou bots (détectés via des patterns récurrents et des métadonnées suspectes)

1.2. Nettoyage du texte

1.2.1. Suppression des caractères spéciaux

Le texte brut contient de nombreux caractères spéciaux : - **URLs** : Suppression des liens raccourcis (format t.co) via expressions régulières - **Caractères non-ASCII** : Normalisation des caractères accentués et gestion des encodages - **Ponctuation excessive** : Réduction des répétitions (ex: "ENCOREEEEE" → "ENCORE") en préservant l'emphase

1.2.2. Gestion des emojis

Les emojis véhiculent des informations sentimentales importantes. La stratégie adoptée : - **Suppression via Regex Unicode** : Les emojis sont supprimés pour ne pas polluer l'analyse

sémantique - **Préservation du contexte émotionnel** : L'information sentimentale est capturée par les modèles LLM lors de l'analyse du texte nettoyé

1.2.3. Traitement des mentions

Les mentions (@username) sont fréquentes dans les tweets SAV : - **Masquage temporaire** : Les mentions sont préservées via un masquage temporaire (__MENTION_i__) durant le nettoyage - **Conservation des mentions pertinentes** : Les mentions de comptes officiels Free (@Free, @Freebox) sont préservées pour le contexte métier

1.3. Ajout de colonnes utiles

Le processus enrichit le dataset avec trois colonnes calculées via des modèles LLM :

1.3.1. Colonne sentiment

Classification automatique du ton émotionnel en trois catégories : - **negative** : Tweets exprimant frustration, colère, mécontentement - **neutral** : Messages factuels, demandes d'information sans charge émotionnelle - **positive** : Expressions de satisfaction, remerciements, retours positifs

Le calcul s'effectue via des modèles LLM (Gemini 1.5 Flash, Mistral Small) analysant le contexte global du message.

1.3.2. Colonne priority

Segmentation en trois niveaux d'urgence basée sur l'analyse sémantique : - **High (2)** : Tweets contenant des indicateurs d'urgence critiques ("injoignable", "depuis 2 jours", "aucun accès", "panne totale") - **Medium (1)** : Demandes nécessitant une réponse rapide mais non critiques - **Low (0)** : Questions générales, demandes d'information sans urgence

La détection combine l'analyse LLM avec des patterns regex pour identifier les expressions d'urgence.

1.3.3. Colonne main_topic

Classification thématique automatique identifiant le sujet principal : - **Réseau** : Problèmes de connexion, pannes, latence - **Facture** : Questions de facturation, tarifs, paiements - **Service Client** : Demandes générales, réclamations administratives - **Équipement** : Problèmes Freebox, matériel - **Abonnement** : Gestion d'abonnement, résiliation, modification

Ainsi, nous aurons un routage automatique vers les services compétents.

1.4. Architecture technique de traitement

La solution Dallosh Analysis repose sur une architecture microservices robuste pour la flexibilité et la supervision en temps réel.

1.4.1. Structure du projet et services clés

Le système est segmenté en trois composants principaux orchestrés via Docker/Docker Compose :

Tableau 1 : Structure du projet et services clés

Composant	**Rôle Principal**	**Services Python Associés**
frontend (Next.js/React)	Interface Utilisateur et Visualisation des KPI	**N/A** (Consommation API REST)
backend (TypeScript/Node.js)	API REST (Auth, Fichiers, Tâches), Base de Données (MongoDB) et Orchestration	**N/A** (Appelle Celery)
microservices/auto_processing_datasets	Cœur du traitement ETL, asynchrone (Celery/RabbitMQ)	cleaning.py, calling_llm.py, saving.py

Rôle des Services Python (Microservices) :

- **src/services/reading_file.py** : Charge le CSV brut en DataFrame Pandas
- **src/services/cleaning.py** : Applique les règles de filtrage (Regex, IQR) pour dédoublonner et assainir le texte
- **src/services/calling_llm.py** : Gère l'appel au LLM pour l'enrichissement (Sentiment, Topic, Priorité) avec logique de retry et gestion de pagination
- **src/tasks/processor.py** : **Orchestrateur Celery**, enchaîne les étapes et gère les statuts de la tâche (TASK_STATUS_IN_QUEUE, TASK_STATUS_DONE, etc.) via RabbitMQ

1.4.2. Architecture de la base de données

Le système utilise MongoDB pour stocker les métadonnées et la configuration.

Tableau 2 : Architecture de la base de données

Collection	Description	Rôle dans l'Analyse
files	Métadonnées de tous les fichiers uploadés (bruts, nettoyés, analysés)	Permet de lier une tâche au chemin du fichier physique sur le stockage
tasks	Historique et statut des pipelines de traitement exécutés	Stocke les étapes complétées, le modèle LLM utilisé et le statut temps réel (RabbitMQ)
settings	Configuration des modèles LLM (APIs, clés, mode local/externe, limites de requêtes)	Garantit l'adaptabilité de la solution aux contraintes externes et aux coûts API
users	Informations des utilisateurs, mot de passe hashé, rôle	Gère l'authentification et l'autorisation
roles	Définition des rôles et des permissions associées	Contrôle l'accès aux fonctionnalités sensibles (e.g., config LLM, vue logs)

1.4.3. Pipeline de nettoyage et structuration

Le cœur du traitement (cleaning.py et calling_llm.py) suit une logique stricte :

1. **Ingestion & Nettoyage :**
   - Suppression des émojis via Regex Unicode pour ne pas polluer l'analyse sémantique
   - Préservation intelligente des mentions (@user) grâce à un masquage temporaire (__MENTION_i__) durant le nettoyage des caractères spéciaux
   - Déduplication stricte et suppression des outliers via la méthode statistique IQR (Interquartile Range)

2. **Enrichissement par IA (Mode Hybride) :**
   - L'application permet de structurer les tweets (Sentiment, Priorité, Topic) via des LLM
   - **Gestion de la stratégie :** Le système supporte un mode "Automatique" avec repli (fallback). Si l'API principale échoue, le système peut basculer sur un modèle local ou secondaire

1.4.4. Supervision et contrôle d’administration

L’interface graphique (React) est le centre de contrôle de la solution. Elle offre aux administrateurs une flexibilité totale sur le pipeline.

Gestion des Utilisateurs et des Rôles (Sécurité)

Le système intègre une gestion complète des utilisateurs et des rôles/permissions. Seuls les administrateurs peuvent configurer les modèles LLM et visualiser les logs techniques. Les analystes ont accès uniquement aux dashboards et à l’upload des datasets.

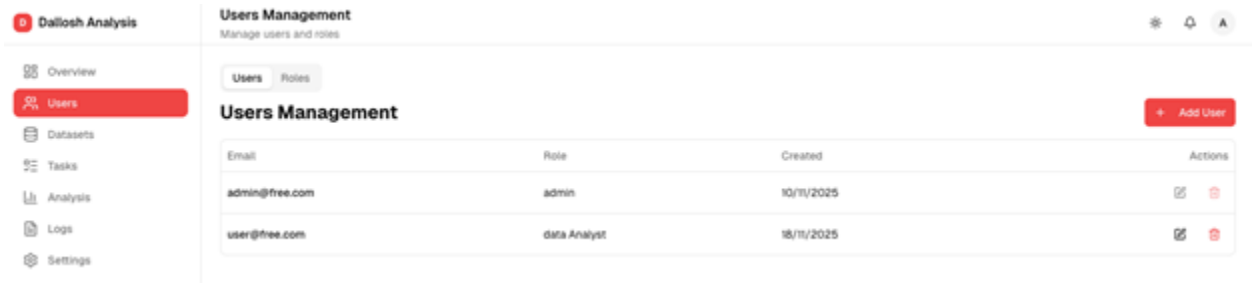

Figure 1 : Gestion des Utilisateurs et Rôles

Orchestration Événementielle (RabbitMQ)

Chaque étape du traitement émet des événements, permettant une supervision granulaire visible dans les logs.

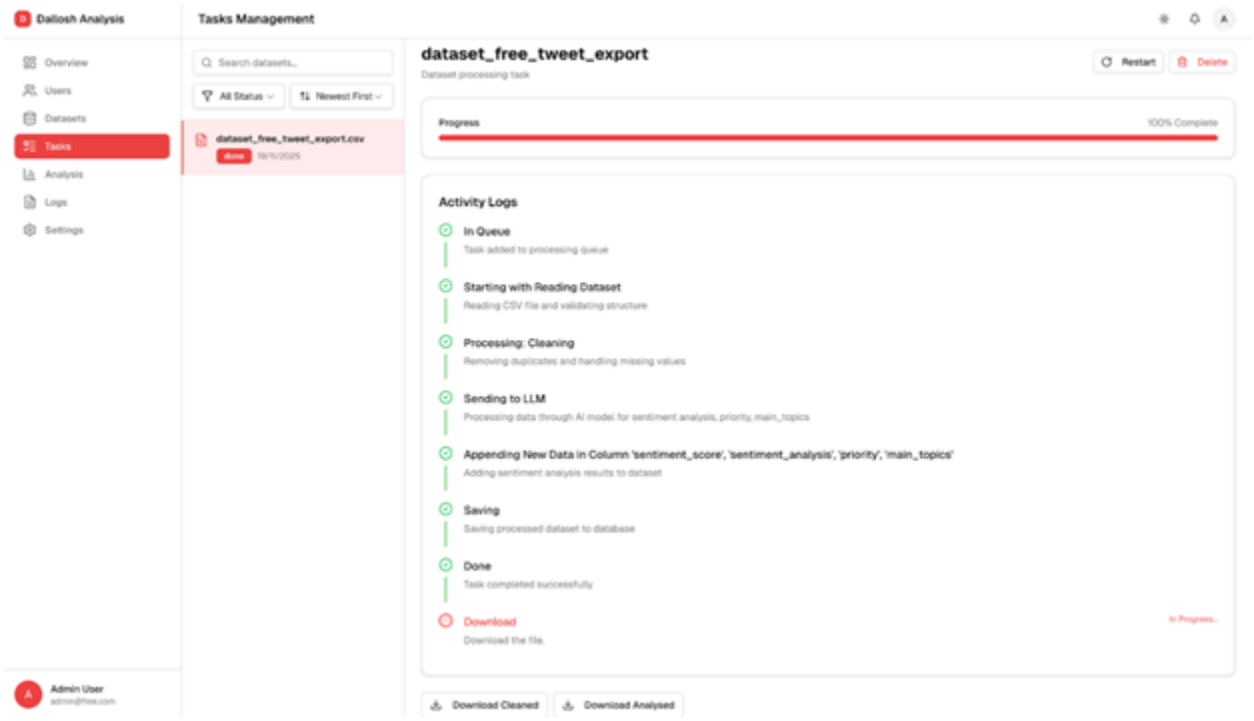

Figure 2 : Vue des Logs Système et Événements

Configuration Modulaire des Modèles IA

Les administrateurs peuvent changer dynamiquement l'API utilisée (Gemini, Mistral, OpenAI) ou opter pour des modèles locaux (Ollama) en cas de coupure réseau ou pour des raisons de confidentialité.

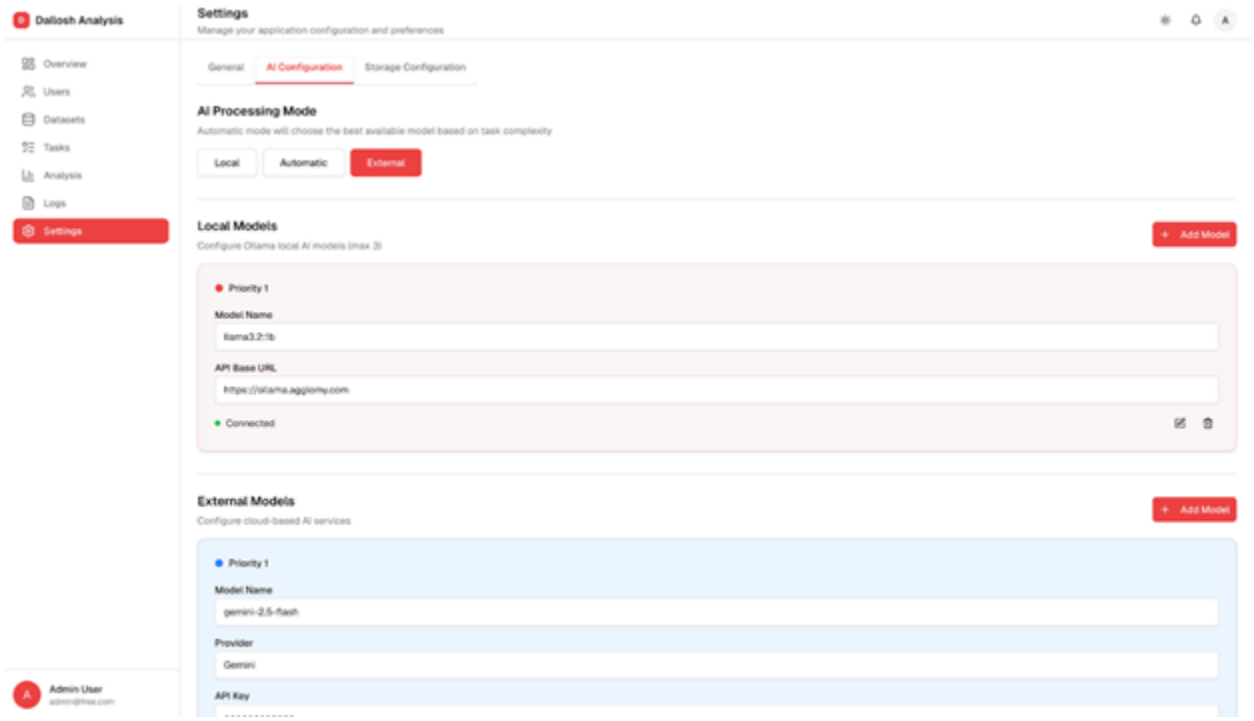

Figure 3 : Interface de Configuration des Modèles IA

2. Définition et calcul des KPI

Les indicateurs sont calculés dynamiquement dans le frontend (DatasetAnalysis.tsx) sur la base du fichier enrichi.

Tableau 3 : Synthèse des KPI

Catégorie	KPI	Définition	Calcul	Métriques Dérivées	Interprétation
Volume et Flux	Volume Net	Nombre de tweets pertinents après filtrage du bruit (retweets, doublons, spam)	Volume Net = Total tweets initiaux - (Retweets + Doublons + Tweets hors sujet)	—	Mesure la charge de travail réelle pour le service client, excluant le bruit
	Ratio de Bruit	Pourcentage de tweets rejetés par rapport au volume initial	Ratio de Bruit = (Tweets rejetés / Total tweets initiaux) × 100	—	Un ratio élevé indique une forte proportion de contenu non pertinent, nécessitant un affinement des critères de filtrage
	Volume de Tweets SAV par Période	Distribution temporelle des tweets pertinents, calculée par jour et par semaine	Agrégation des tweets par date (created_at) après filtrage, avec regroupement journalier et hebdomadaire	—	Permet d'identifier les pics d'activité, les tendances saisonnières et de dimensionner les ressources
Satisfaction et Urgence	Répartition des Sentiments	Distribution des tweets selon leur tonalité émotionnelle (positive, neutre, négative)	Comptage des occurrences de chaque valeur de la colonne sentiment, exprimé en nombre absolu et en pourcentage	**NSS :** (Tweets positifs - Tweets négatifs) / Total tweets × 100**Taux de Négativité :** Tweets négatifs / Total tweets × 100	Un NSS négatif structurel est attendu en SAV, mais les dérives importantes signalent des problèmes systémiques nécessitant une intervention
	Indice de Priorité	Segmentation des tweets selon leur niveau d'urgence (High, Medium, Low)	Distribution des valeurs de la colonne priority avec focus sur les tweets High (priorité 2)	**Taux de Priorité Haute :** Tweets High / Total tweets × 100**Volume de Crise :** Nombre absolu de tweets nécessitant une intervention < 1h	Les tweets High Priority nécessitent un traitement immédiat. Un volume élevé signale une situation de crise nécessitant une mobilisation exceptionnelle

	Cartographie des Thèmes	Répartition des tweets selon leur sujet principal identifié (main_topic)	Comptage des occurrences de chaque thème, avec analyse croisée avec le sentiment	**Thème le plus fréquent :** Identification du sujet dominant**Thème le plus négatif :** Croisement main_topic × sentiment	Permet d'identifier les domaines nécessitant une attention particulière et d'optimiser le routage
Engagement	Temps Moyen de Réponse	Délai moyen entre le tweet initial du client et la première réponse de Free (si applicable)	Temps Moyen = Moyenne(différence entre created_at du tweet et created_at de la réponse)	—	Indicateur de réactivité du service client. Un temps élevé peut expliquer l'escalade des plaintes
	Taux de Réponse (Reply Rate)	Proportion de tweets ayant reçu une réponse officielle de Free	Reply Rate = (Tweets avec in_reply_to non null / Total tweets) × 100	—	Mesure le niveau d'interactivité et d'engagement du service client sur Twitter
	Viralité des Plaintes	Corrélation entre le sentiment négatif et l'engagement (retweets, likes)	Analyse de corrélation entre colonne sentiment (négatif) et colonnes retweet_count et favorite_count	**Score de Viralité :** (retweet_count + favorite_count) pour tweets négatifs / Total engagement**Amplification des Plaintes :** Ratio engagement tweets négatifs / engagement tweets positifs	Les plaintes virales peuvent amplifier l'impact négatif et nécessitent une réponse rapide

3. Exploration visuelle

L'application génère des tableaux de bord interactifs permettant à la direction de Free de visualiser et interpréter les données.

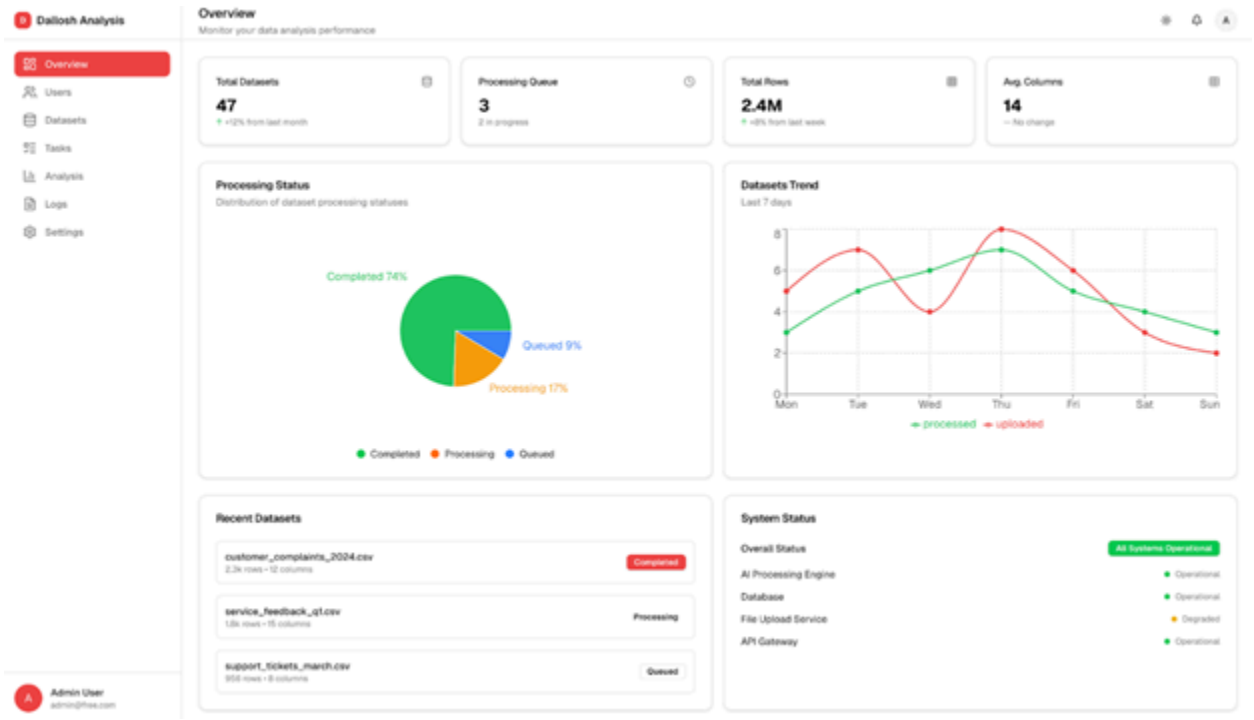

Figure 4 : Tableau de Bord Analytique Global

3.1. Histogrammes de Volume

3.1.1. Distribution Temporelle

Visualisation : Histogramme montrant le volume de tweets par jour/semaine sur la période analysée.

Interprétation : Permet d'identifier les pics d'activité, les tendances saisonnières ou hebdomadaires, et les corrélations avec des événements externes (pannes réseau, campagnes marketing).

3.1.2. Répartition par Thèmes

Visualisation : Histogramme horizontal ou vertical montrant la distribution des tweets par main_topic.

Interprétation : Identifie les sujets les plus fréquents, permettant de prioriser les ressources et les formations.

3.2. Courbes d’Évolution Temporelle

3.2.1. Évolution du Sentiment dans le Temps

Visualisation : Courbes temporelles superposées montrant l’évolution du volume de tweets positifs, neutres et négatifs.

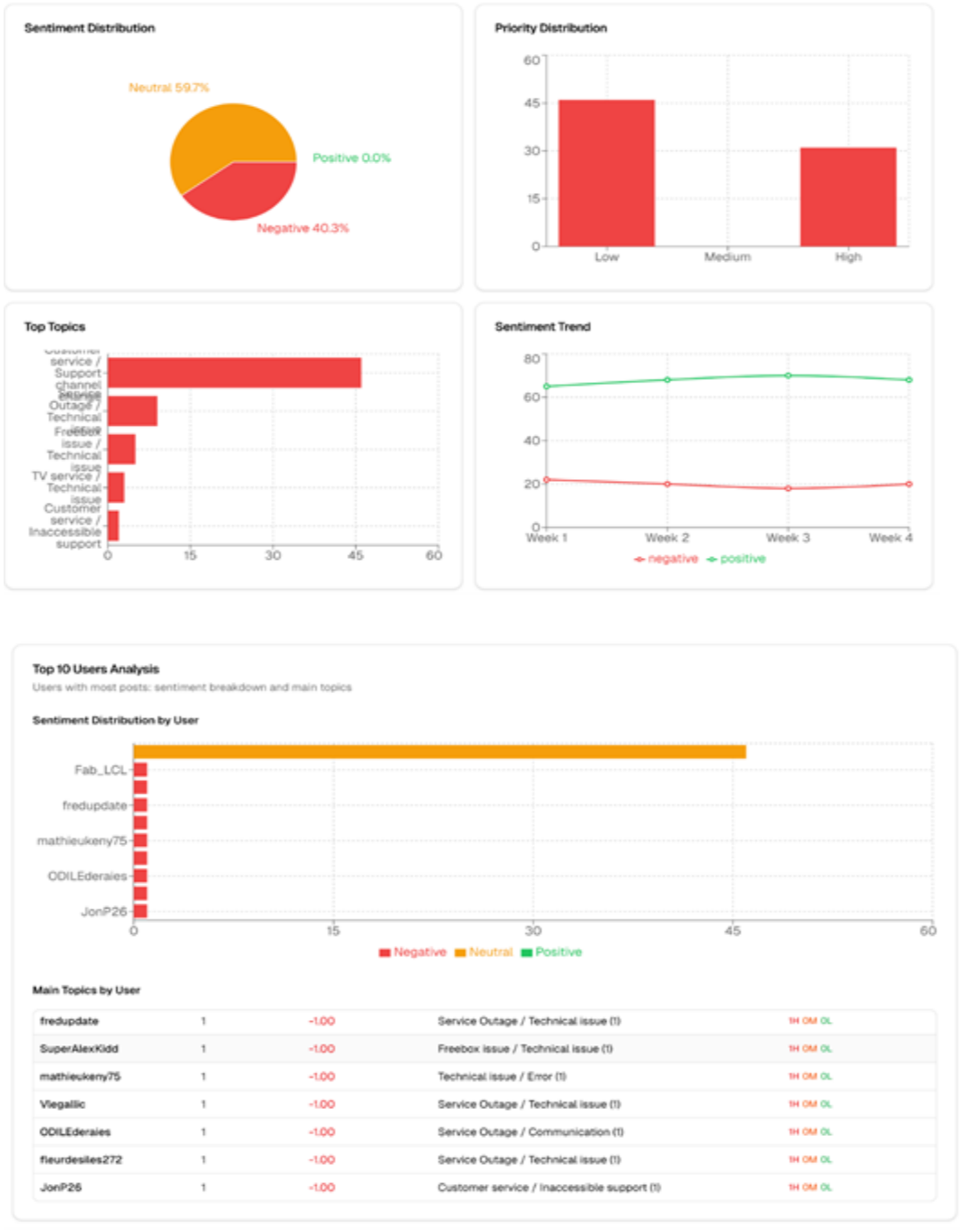

Figure 5 : Timeline de l’Évolution des Sentiments

Interprétation :

- **Détection d'incidents majeurs :** Une hausse concomitante du volume et du sentiment négatif signale une panne réseau ou un problème systémique
- **Tendances à long terme :** Permet d'évaluer l'impact des actions correctives sur la satisfaction client
- **Signaux faibles :** Une lente montée des plaintes sur plusieurs jours permet de prévenir l'ingénierie avant la saturation du SAV

3.2.2. Évolution des Priorités

Visualisation : Courbe temporelle montrant l'évolution du nombre de tweets High Priority.

Interprétation : Permet d'anticiper les situations de crise et de mobiliser les ressources.

3.3. Nuages de Mots

3.3.1. Nuage de Mots pour Tweets Négatifs

Visualisation : Nuage de mots généré à partir du texte des tweets classés comme négatifs, avec taille proportionnelle à la fréquence.

Interprétation : Identifie les mots-clés récurrents dans les plaintes, détecte les problèmes récurrents (ex: "coupure", "panne", "lenteur") et permet de comprendre les préoccupations principales des clients mécontents.

Méthodologie : Extraction des mots significatifs (exclusion des stopwords), pondération par fréquence et importance sémantique, visualisation avec taille et couleur variables.

3.4. Cartographie Thématique

3.4.1. Distribution des Topics et Priorités

Visualisation : Histogramme empilé croisant les thèmes (main_topic) et les niveaux de priorité, avec code couleur pour les sentiments.

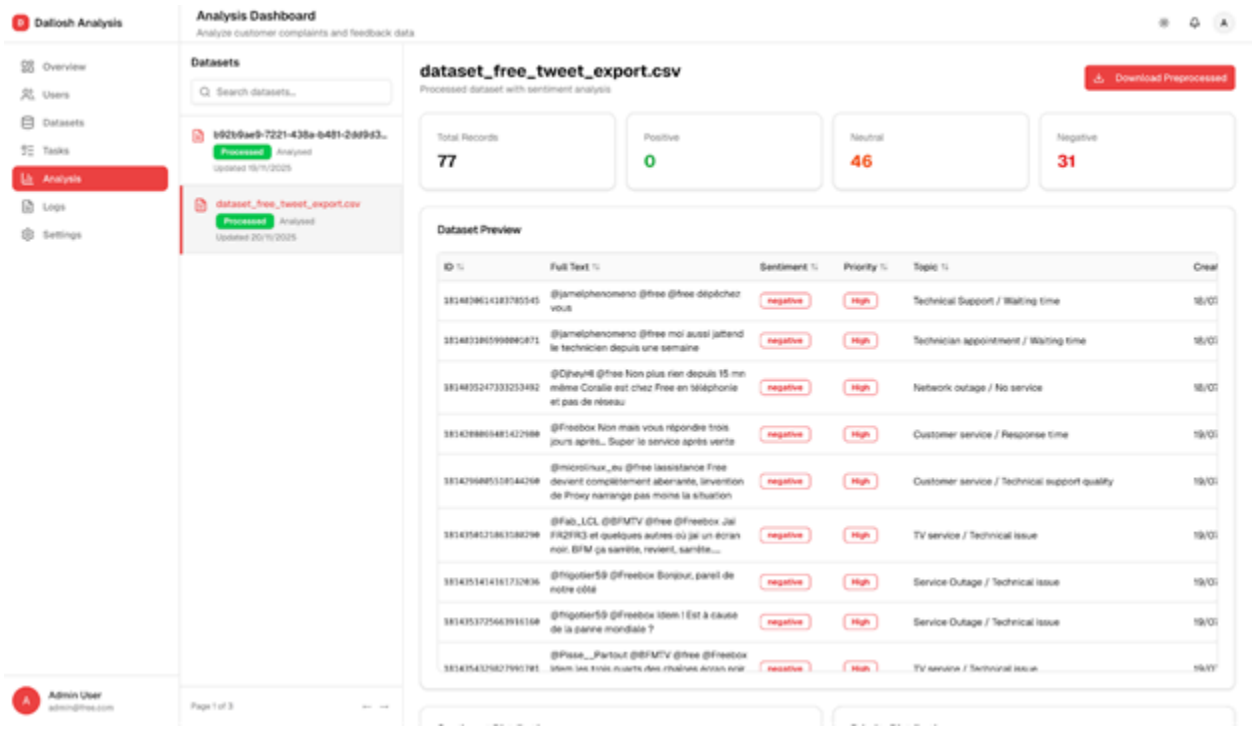

Figure 6 : Distribution des Topics et Priorités

Interprétation : - **Constat principal :** Le sujet “Réseau” concentre le plus fort taux de négativité et de priorité haute, signalant un domaine critique nécessitant une attention particulière - **Sujets administratifs :** Majoritairement neutres, indiquant des demandes d’information plutôt que des plaintes - **Zones à risque :** Combinaisons thème × priorité nécessitant une intervention urgente

4. Qualité des données

4.1. Proportion de Tweets Rejetés

L’analyse de la qualité montre que le processus de nettoyage est efficace mais drastique.

4.1.1. Taux de Rejet Global

Calcul : Proportion de tweets éliminés lors des différentes étapes de filtrage.

Composition du rejet : - **Retweets :** Proportion variable selon la période (estimée entre 15-25% du dataset initial) - **Doublons :** Environ 5-10% du dataset (tweets quasi-identiques) - **Tweets hors sujet :** 10-15% (spam, humour, messages promotionnels)

Impact : Le volume net final représente environ 60-70% du dataset initial, ce qui est cohérent avec les attentes pour un dataset de tweets bruts.

4.1.2. Bénéfices et Limites du Filtrage

Bénéfices : Élimination du bruit pour une analyse plus précise, concentration sur les interactions directes avec Free, métriques calculées sur des données pertinentes.

Limites : L'élimination des retweets peut masquer l'ampleur d'une campagne de dénigrement coordonnée. Certains retweets avec commentaires ajoutés peuvent contenir des informations pertinentes. La frontière entre "hors sujet" et "pertinent" peut être floue.

4.2. Limites du Jeu de Données

4.2.1. Manques et Données Incomplètes

Métadonnées manquantes : Certains tweets ne contiennent pas d'informations de géolocalisation. Les métadonnées d'engagement (views_count) peuvent être incomplètes pour les tweets anciens. Les threads de discussion ne sont pas toujours reconstituables.

Impact : Certaines analyses approfondies (analyse géographique, reconstruction de conversations) sont limitées.

4.2.2. Bruit et Ambiguïté

Bruit résiduel : Certains tweets utilisent un langage tellement familier ou abrégé que même les LLM peinent à extraire le sens. La détection imparfaite de l'ironie et du sarcasme peut conduire à des classifications erronées. Un tweet peut contenir plusieurs demandes distinctes, compliquant la classification unique.

Ambiguïté contextuelle : Tweets courts faisant référence à des situations non explicitement mentionnées. Certains tweets nécessitent la connaissance d'événements externes (pannes, annonces) pour être correctement interprétés.

4.2.3. Limites Techniques du Traitement

Traitement par lots fixes : Le traitement actuel par lots fixes (500 lignes) ne tient pas compte de la longueur variable des textes. Si la limite de tokens du LLM est atteinte, des tweets longs peuvent être tronqués, perdant des informations importantes. Optimisation future : passage à une logique de "tokens" plutôt que de "lignes" pour maximiser l'utilisation des quotas API.

Variabilité des performances LLM : Les modèles peuvent produire des résultats variables selon la complexité du tweet. Nécessité de validation manuelle sur un échantillon pour garantir la qualité.

4.3. Recommandations pour l'Amélioration de la Qualité

4.3.1. Affinement des Critères de Filtrage

Révision périodique des critères de "hors sujet" basée sur le retour d'expérience des agents. Conservation sélective des retweets avec commentaires ajoutés significatifs.

4.3.2. Enrichissement des Métadonnées

Développement d'algorithmes pour reconstituer les threads de conversation. Extraction et normalisation des informations de localisation quand disponibles.

4.3.3. Validation et Feedback

Intégration des corrections manuelles des agents pour améliorer les modèles. Calcul de métriques de confiance pour chaque classification (score de confiance LLM).

Conclusion

Ce rapport présente une méthodologie complète de nettoyage, structuration et exploration des tweets SAV Free. Le processus de filtrage isole environ 60-70% de tweets pertinents, enrichis avec des colonnes sémantiques (sentiment, priorité, thème) calculées via des modèles LLM.

Les KPI définis offrent une vision multidimensionnelle de l'activité SAV : volume, satisfaction, urgence, engagement. Les visualisations interactives permettent à la direction de Free de détecter les incidents majeurs, d'identifier les tendances et d'anticiper les situations de crise.

La solution Dallosh offre une infrastructure résiliente (RabbitMQ) et flexible (choix des modèles LLM), capable de transformer des données brutes en insights visuels exploitables pour la prise de décision stratégique.

Limites identifiées : Le processus de nettoyage, bien qu'efficace, peut masquer certains phénomènes (bad buzz artificiel). Les limites techniques actuelles (traitement par lots fixes) nécessitent des améliorations futures (optimisation par tokens, validation manuelle renforcée).

Perspectives : L'intégration d'une boucle de feedback avec les agents permettra d'améliorer continuellement la qualité des classifications et de raffiner les critères de filtrage.