



RAPPORT PROJET FIN MODULE

Analyse et prédition des marchés boursiers à l'aide des données Big Data et du sentiment des réseaux sociaux

REALISE PAR :

SALMA HMITTI
MOHAMED ENAHARI
YOUSSEF MOUNTASSIR
YOUSSEF EL HARDOU

ENCARDRE PAR :

PR. YASSER EL MADANI EL AMALI

Année Universitaire : 2025/2026

REMERCIEMENTS

On tient à remercier dans un premier temps, toute l'équipe pédagogique de l'**ISMAGI** Rabat et les intervenants professionnels responsables de la formation.

Et avant d'entamer ce rapport, nous profitons de l'occasion pour remercier tout d'abord notre professeur **Monsieur YASSER EL MADANI EL ALAMI** qui n'a pas cessé de nous encourager pendant la durée du projet, ainsi pour sa générosité en matière de big data NoSql.

Nous la remercions également pour l'aide et les conseils concernant les missions évoquées dans ce Projet , qu'elle nous a apporté lors des différents suivis, et la confiance qu'elle nous a témoignée.

Avec toute notre reconnaissance, on vous prie d'agrérer, **Monsieur** , l'expression de nos salutations distinguées.

Résumé

Ce projet vise à concevoir et implémenter un **système de prédition boursière en temps réel** reposant sur l'analyse combinée des **données financières** et du **sentiment exprimé sur les réseaux sociaux**, en particulier Reddit à travers le subreddit *WallStreetBets*.

L'architecture proposée exploite des technologies Big Data modernes telles que **Apache Kafka**, **Apache Spark**, **MongoDB** et **Docker**, permettant l'ingestion, le traitement et le stockage de flux de données volumineux et hétérogènes. Les données textuelles issues de Reddit sont analysées afin d'extraire des indicateurs de sentiment, qui sont ensuite fusionnés avec les données boursières provenant de **Stooq**. Des modèles de **Deep Learning de type LSTM** sont utilisés pour prédire les mouvements futurs des actions.

Les résultats sont visualisés à travers un **tableau de bord interactif Streamlit**, offrant une vision globale et dynamique des tendances du marché et de l'activité sociale associée.

Abstract

This project aims to design and implement a **real-time stock market prediction system** based on the combined analysis of **financial data** and **social media sentiment**, particularly from Reddit's *WallStreetBets* community.

The proposed system relies on a modern **Big Data architecture** using **Apache Kafka**, **Apache Spark**, **MongoDB**, and **Docker** to handle large-scale, heterogeneous data streams.

Reddit textual data are processed to extract sentiment indicators, which are then merged with stock market data obtained from **Stooq**. **LSTM-based deep learning models** are applied to predict future stock movements.

The results are presented through an **interactive Streamlit dashboard**, providing real-time insights into both market trends and social sentiment dynamics.

Liste des abréviations

- **API** : Application Programming Interface
- **ML** : Machine Learning
- **DL** : Deep Learning
- **XGBoost** : eXtreme Gradient Boosting
- **KPI** : Key Performance Indicator
- **ETL** : Extract, Transform, Load
- **DAG** : Directed Acyclic Graph

Liste des figures

Figure 1 : Architecture globale du système Big Data

Figure 2 : Orchestration des services Big Data avec Docker Compose

Figure 3 : Pipeline de traitement des données

Figure 4 : Pipeline Spark pour le traitement des données Reddit

Figure 5 : Python logo

Figure 6 : Apache kafka logo

Figure 7 : Apach Spark Structured Streaming

Figure 8 : MongoDB .BD NosSQL

Figure 9 : docker logo

Figure 10 : docker compose

Figure 11 : Vue d'ensemble du tableau de bord

Figure 12 : Évolution du prix de l'action

Figure 13 : Analyse du volume de transactions

Figure 14 : Fréquence des publications Reddit

Figure 15 : Score de sentiment Reddit

Figure 16 : Comparaison prix réel vs prix prédit & Erreur de prédiction dans le temps

SOMMAIRE

REMERCIEMENTS	2
<i>Résumé</i>	3
Abstract.....	4
Liste des abréviations.....	5
• XGBoost : eXtreme Gradient Boosting	5
Liste des figures	6
SOMMAIRE	7
Introduction générale	11
• Contexte et enjeux des marchés financiers	11
• Impact des réseaux sociaux sur la bourse	11
• Problématique du projet.....	11
• Objectifs généraux	12
• Organisation du rapport.....	12
Chapitre 1 : Présentation du projet	13
1.1 Description globale du système	13
1.2 Cas d'usage du système	13
1.3 Valeur ajoutée du projet.....	14
1.4 Positionnement Big Data et Intelligence Artificielle	14
Chapitre 2 : Étude de l'existant.....	15
2.1 Analyse des approches classiques de prédiction boursière	15
2.2 Utilisation du Machine Learning dans la prédiction boursière	15
2.3 Analyse de l'utilisation du sentiment analysis en finance.....	16
2.4 Limites des solutions existantes.....	16
2.5 Justification de la solution proposée.....	16

Chapitre 3 : Sources de données	18
3.1 Importance des données dans un système de prédition boursière	18
3.2 Données issues de Reddit (WallStreetBets – Kaggle)	18
3.3 Données boursières (Stooq).....	19
3.4 Fusion et cohérence des sources de données.....	20
3.5 Justification du choix des sources de données	21
Chapitre 4 : Architecture du système Big Data	22
4.1 Vue d'ensemble de l'architecture.....	22
4.2 Couche des sources de données	24
4.3 Couche de collecte des données.....	24
4.4 Couche de streaming – Apache Kafka	24
4.5 Couche de traitement – Apache Spark Streaming.....	25
4.6 Couche de stockage – MongoDB.....	25
4.7 Couche Machine Learning	25
4.8 Couche d'orchestration – Apache Airflow	26
4.9 Couche de visualisation	26
4.10 Conteneurisation et déploiement – Docker	26
Conclusion du chapitre	28
Chapitre 5 : Pipeline de traitement des données	29
5.1 Objectifs du pipeline de traitement	29
5.2 Ingestion des données en temps réel	29
5.3 Nettoyage et prétraitement des données	30
5.4 Analyse de sentiment des posts Reddit.....	30
5.5 Agrégation temporelle des données	31
5.6 Fusion des données sociales et financières.....	31
5.7 Feature engineering.....	31

5.8 Stockage des données traitées.....	32
5.9 Gestion des erreurs et robustesse du pipeline	32
5.10 Résumé du pipeline de traitement.....	32
Conclusion du chapitre	33
Chapitre 6 : Modélisation et Machine Learning.....	34
6.1 Objectifs de la modélisation	34
6.2 Choix des modèles de prédiction	34
6.3 Présentation du modèle XGBOOT	34
6.4 Données d'entrée du modèle.....	35
6.5 Préparation des données pour l'apprentissage	35
6.6 Architecture du modèle XGBOOT.....	36
6.7 Entraînement du modèle	36
6.8 Modèles complémentaires et comparaison	37
6.9 Métriques d'évaluation.....	37
6.10 Suivi des expériences avec MLflow	37
6.11 Résumé de la phase de modélisation.....	38
Conclusion du chapitre	38
Chapitre 7 – Implémentation technique.....	39
7.1 Introduction.....	39
7.2 Choix technologiques.....	39
7.3 Déploiement et orchestration.....	40
7.4 Intégration des composants.....	40
Conclusion du chapitre	40
Chapitre 8 : Visualisation et tableau de bord.....	41
8.1 Objectifs de la visualisation.....	41
8.2 Outil de visualisation	41

8.3 Indicateurs clés affichés (KPI)	41
8.4 Visualisation des données boursières	41
8.5 Visualisation de l'activité et du sentiment Reddit	44
8.6 Visualisation des prédictions.....	45
8.7 Conclusion du chapitre	46
Chapitre 9 : Résultats et analyse	47
9.1 Résultats des prédictions	47
9.2 Analyse de l'impact du sentiment Reddit.....	47
9.3 Étude de cas (GME / AMC / TSLA).....	47
9.4 Discussion des performances.....	47
Conclusion du chapitre	47
Chapitre 10 : Limites du projet	48
10.1 Limites techniques.....	48
10.2 Limites liées aux données	48
10.3 Limites des modèles	48
Conclusion du chapitre	48
Chapitre 11 : Perspectives et améliorations	49
11.1 Enrichissement des sources de données.....	49
11.2 Amélioration des modèles de prédiction	49
11.3 Déploiement et passage à l'échelle	49
11.4 Fonctionnalités avancées.....	49
Conclusion du chapitre	49
Conclusion générale	50
Références bibliographiques	51
Résumé global du rapport.....	52

Introduction générale

• Contexte et enjeux des marchés financiers

Les marchés financiers occupent une place centrale dans l'économie mondiale, en permettant la circulation des capitaux entre les investisseurs et les entreprises. Ils jouent un rôle fondamental dans le financement de l'activité économique, l'évaluation de la valeur des entreprises et la gestion des risques financiers. Cependant, ces marchés sont caractérisés par une forte volatilité, rendant les décisions d'investissement complexes et incertaines.

Traditionnellement, l'analyse des marchés financiers repose sur des indicateurs économiques et financiers classiques tels que les prix historiques, les volumes d'échange, les ratios financiers ou encore les annonces macroéconomiques. Bien que ces indicateurs restent essentiels, ils ne suffisent plus à expliquer l'ensemble des fluctuations observées sur les marchés. En effet, les comportements des investisseurs sont également influencés par des facteurs psychologiques, émotionnels et sociaux, qui peuvent provoquer des variations rapides et parfois imprévisibles des cours boursiers.

Dans ce contexte, l'exploitation de nouvelles sources de données et l'utilisation de technologies avancées deviennent indispensables pour mieux comprendre et anticiper les dynamiques des marchés financiers.

• Impact des réseaux sociaux sur la bourse

Avec l'essor du numérique, les réseaux sociaux sont devenus des espaces majeurs d'échange d'informations et d'opinions. Des plateformes telles que Reddit, Twitter ou encore Facebook permettent à des millions d'utilisateurs de partager en temps réel leurs analyses, leurs émotions et leurs décisions d'investissement. Ces interactions collectives contribuent à façonner le sentiment général du marché, souvent appelé *sentiment des investisseurs*.

Reddit, et plus particulièrement le subreddit WallStreetBets, illustre parfaitement ce phénomène. Cette communauté a démontré sa capacité à influencer significativement les cours de certaines actions, comme lors de l'affaire GameStop, où une mobilisation massive d'investisseurs particuliers a provoqué des mouvements de marché exceptionnels. Cet événement a mis en évidence le pouvoir des réseaux sociaux en tant que catalyseurs de tendances boursières.

Ainsi, les réseaux sociaux ne sont plus de simples plateformes de discussion, mais constituent de véritables sources d'information stratégique, capables d'anticiper ou d'amplifier les mouvements du marché. Leur analyse représente donc un enjeu majeur pour la finance moderne.

• Problématique du projet

Malgré l'abondance des données disponibles sur les réseaux sociaux et les marchés financiers, leur exploitation efficace reste un défi majeur. Ces données sont caractérisées par un volume important,

une grande variété (textes, chiffres, séries temporelles) et une vitesse élevée, notamment dans un contexte de diffusion en temps réel.

La problématique centrale de ce projet est donc la suivante :

Comment exploiter efficacement les données issues des réseaux sociaux, combinées aux données financières, afin de prédire les mouvements boursiers en temps réel dans un contexte Big Data ?

Cette problématique soulève plusieurs défis techniques et méthodologiques, notamment la gestion du streaming de données, l'analyse automatique du langage naturel, la modélisation de séries temporelles complexes et la visualisation dynamique des résultats.

• Objectifs généraux

Afin de répondre à cette problématique, ce projet poursuit plusieurs objectifs complémentaires :

- Collecter et traiter des données financières et sociales en temps réel, en s'appuyant sur des sources fiables telles que Stooq pour les cours boursiers et Reddit pour les discussions financières.
- Analyser le sentiment des discussions Reddit, à travers des techniques de traitement automatique du langage naturel, afin d'extraire des indicateurs reflétant l'opinion des investisseurs.
- Concevoir et entraîner un modèle de prédiction basé sur le Deep Learning, notamment des réseaux de neurones récurrents de type LSTM, capables de capturer les dépendances temporelles des données financières.
- Visualiser les résultats de manière interactive, à l'aide d'un tableau de bord dynamique permettant de suivre l'évolution des marchés et du sentiment social en temps réel.

Ces objectifs visent à démontrer l'apport des architectures Big Data et de l'intelligence artificielle dans le domaine de la finance prédictive.

• Organisation du rapport

Ce rapport est structuré de manière progressive afin de guider le lecteur à travers les différentes étapes du projet. Il débute par une présentation générale du système et une étude de l'existant, avant de détailler les sources de données utilisées. Les chapitres suivants sont consacrés à l'architecture Big Data, au pipeline de traitement des données et à la modélisation par Machine Learning.

Enfin, le rapport présente les résultats obtenus, analyse les performances du système, discute les limites rencontrées et propose des perspectives d'amélioration. Cette organisation permet d'offrir une vision complète, cohérente et professionnelle du travail réalisé.

Chapitre 1 : Présentation du projet

1.1 Description globale du système

Le projet consiste à concevoir un **système intelligent de prédition boursière en temps réel** basé sur l'exploitation conjointe des **données financières** et du **sentiment exprimé sur les réseaux sociaux**, en particulier Reddit via la communauté *WallStreetBets*.

L'objectif du système est d'analyser en continu ces différentes sources d'information afin de **détecter des tendances, anticiper les mouvements de prix et fournir des indicateurs décisionnels exploitables**.

Le système repose sur une **architecture Big Data distribuée**, capable de gérer :

- des flux de données en streaming,
- de grands volumes de données historiques,
- des données hétérogènes (textes et séries temporelles).

Les données Reddit sont traitées afin d'extraire des **scores de sentiment**, tandis que les données boursières sont utilisées pour représenter l'évolution réelle des marchés. Ces deux types d'informations sont ensuite fusionnés et exploités par des **modèles de Deep Learning** pour produire des prédictions.

1.2 Cas d'usage du système

Le système est principalement appliqué à l'analyse d'actions fortement discutées sur Reddit, notamment :

- **GameStop (GME)**
- **AMC Entertainment (AMC)**
- **Tesla (TSLA)**

Ces actions sont connues pour leur forte volatilité et leur sensibilité au sentiment des investisseurs particuliers. Le cas d'usage typique du système est le suivant :

1. Les utilisateurs publient des messages et commentaires sur Reddit concernant une action donnée.
2. Le système collecte ces messages en temps réel et analyse leur contenu textuel.
3. Un score de sentiment est calculé afin de déterminer si l'opinion générale est positive, négative ou neutre.
4. En parallèle, les cours boursiers de l'action sont collectés.
5. Le modèle de prédition exploite ces informations pour anticiper les mouvements futurs du prix.
6. Les résultats sont affichés dans un tableau de bord interactif.

Ce cas d'usage illustre la capacité du système à **relier l'activité sociale aux dynamiques du marché financier**.

1.3 Valeur ajoutée du projet

La valeur ajoutée principale de ce projet réside dans l'intégration du **sentiment des investisseurs** comme variable prédictive complémentaire aux indicateurs financiers traditionnels. Contrairement aux approches classiques basées uniquement sur les séries temporelles de prix, ce projet :

- exploite des **données sociales en temps réel**,
- combine analyse de sentiment et prédition boursière,
- utilise une **architecture Big Data scalable**,
- propose une **visualisation interactive** facilitant l'interprétation des résultats.

Cette approche permet d'obtenir une vision plus globale et plus réactive du marché, en tenant compte à la fois des données objectives (prix, volume) et des données subjectives (opinion des investisseurs).

1.4 Positionnement Big Data et Intelligence Artificielle

Le projet se positionne clairement à l'intersection de trois domaines majeurs :

- **Big Data**, à travers la gestion de flux de données volumineux, rapides et variés (Kafka, Spark, MongoDB) ;
- **Intelligence Artificielle**, via l'utilisation de modèles de Deep Learning capables d'apprendre des relations complexes ;
- **Finance prédictive**, en appliquant ces technologies à l'anticipation des mouvements boursiers.

Ce positionnement multidisciplinaire confère au projet une dimension à la fois **académique et professionnelle**, répondant aux enjeux actuels de la finance moderne.

Chapitre 2 : Étude de l'existant

2.1 Analyse des approches classiques de prédition boursière

La prédition des marchés financiers est un domaine étudié depuis plusieurs décennies. Les approches classiques reposent principalement sur l'analyse des **séries temporelles financières** et se divisent en deux grandes catégories : l'analyse fondamentale et l'analyse technique.

L'analyse fondamentale s'appuie sur l'étude des indicateurs économiques et financiers des entreprises, tels que les résultats financiers, les ratios de performance, les annonces macroéconomiques ou encore les politiques monétaires. Bien que pertinente pour des stratégies d'investissement à long terme, cette approche présente des limites importantes dans un contexte de **prédition à court terme** et de **réactivité en temps réel**.

L'analyse technique, quant à elle, repose sur l'exploitation des données historiques des prix et des volumes afin d'identifier des tendances ou des motifs récurrents. Des indicateurs tels que les moyennes mobiles, le RSI (Relative Strength Index) ou le MACD (Moving Average Convergence Divergence) sont couramment utilisés. Toutefois, ces méthodes supposent que les comportements passés se reproduisent, ce qui n'est pas toujours le cas dans des marchés fortement influencés par des événements externes imprévus.

Ces approches classiques, bien qu'utiles, restent **insuffisantes pour capturer la complexité et la volatilité actuelles des marchés financiers**, en particulier dans un environnement dominé par l'information instantanée.

2.2 Utilisation du Machine Learning dans la prédition boursière

Avec l'augmentation de la puissance de calcul et la disponibilité massive de données, le **Machine Learning** s'est imposé comme une alternative prometteuse aux méthodes traditionnelles. Des modèles tels que la régression linéaire, les arbres de décision, les forêts aléatoires ou encore les réseaux de neurones artificiels ont été appliqués à la prédition des cours boursiers.

Ces techniques permettent d'identifier des relations complexes et non linéaires entre différentes variables financières. Cependant, les modèles classiques de Machine Learning restent limités lorsqu'il s'agit de traiter des **données séquentielles**, comme les séries temporelles, car ils ne prennent pas toujours en compte la dépendance temporelle entre les observations.

Pour pallier cette limitation, des modèles plus avancés, tels que les réseaux de neurones récurrents (RNN), ont été introduits. Parmi eux, les **Long Short-Term Memory (LSTM)** se distinguent par leur capacité à mémoriser des informations sur de longues périodes, ce qui les rend particulièrement adaptés à la modélisation des séries financières.

2.3 Analyse de l'utilisation du sentiment analysis en finance

Le **sentiment analysis**, ou analyse de sentiment, vise à extraire l'opinion ou l'émotion exprimée dans un texte. Dans le domaine financier, cette technique est utilisée pour mesurer le sentiment des investisseurs à partir de sources textuelles telles que les articles de presse, les forums ou les réseaux sociaux.

Plusieurs études ont montré que le sentiment des investisseurs peut influencer significativement les marchés boursiers, notamment à court terme. Les plateformes sociales comme Twitter et Reddit permettent d'accéder à un flux continu d'opinions, offrant ainsi une source d'information complémentaire aux données financières traditionnelles.

Toutefois, l'analyse de sentiment présente plusieurs défis, notamment :

- le bruit et l'ambiguïté du langage naturel,
- l'ironie et le sarcasme fréquents sur les réseaux sociaux,
- le déséquilibre entre les opinions positives et négatives.

Malgré ces difficultés, l'intégration du sentiment analysis dans les modèles de prédiction boursière a démontré un **potentiel significatif d'amélioration des performances**, en particulier lorsqu'elle est combinée à des données quantitatives.

2.4 Limites des solutions existantes

Malgré les avancées réalisées, les solutions existantes présentent plusieurs limites importantes :

- **Absence de traitement en temps réel** : de nombreux travaux se concentrent sur l'analyse de données historiques sans intégrer de flux de données continus.
- **Difficulté à gérer le Big Data** : les architectures traditionnelles ne sont pas conçues pour traiter des volumes massifs de données hétérogènes en streaming.
- **Faible intégration des données sociales** : certaines approches exploitent les réseaux sociaux de manière isolée, sans réelle fusion avec les données financières.
- **Manque de visualisation décisionnelle** : peu de solutions proposent des tableaux de bord interactifs permettant une interprétation claire des résultats.

Ces limites soulignent la nécessité d'une approche plus globale, intégrant à la fois les aspects Big Data, Machine Learning et visualisation.

2.5 Justification de la solution proposée

Face aux limites identifiées, ce projet propose une solution innovante reposant sur une **architecture Big Data complète et intégrée**. L'utilisation d'outils tels que **Kafka** et **Spark** permet de traiter des flux de données en temps réel, tandis que **MongoDB** offre une solution flexible pour le stockage des données hétérogènes.

Le recours à des **modèles de Deep Learning de type LSTM** permet de capturer les dépendances temporelles complexes des séries financières, tandis que l'intégration du **sentiment analysis** enrichit les données d'entrée par une dimension comportementale essentielle.

Enfin, le développement d'un **tableau de bord interactif** permet de rendre les résultats accessibles et exploitables, aussi bien dans un cadre académique que professionnel. Cette solution répond ainsi aux enjeux actuels de la finance prédictive et s'inscrit pleinement dans une démarche Big Data moderne.

Chapitre 3 : Sources de données

3.1 Importance des données dans un système de prédiction boursière

La performance d'un système de prédiction boursière repose en grande partie sur la **qualité, la diversité et la pertinence des données utilisées**. Dans un contexte de marchés financiers influencés à la fois par des facteurs économiques et comportementaux, il devient essentiel de combiner des **données quantitatives** (cours boursiers) et des **données qualitatives** (opinions et sentiments des investisseurs).

Dans ce projet, deux sources de données complémentaires ont été exploitées :

- les **données issues des réseaux sociaux**, en particulier Reddit, afin de capturer le sentiment collectif des investisseurs ;
- les **données boursières**, représentant l'évolution réelle des prix et des volumes sur les marchés financiers.

Cette combinaison permet d'obtenir une vision plus complète des forces influençant les mouvements boursiers.

3.2 Données issues de Reddit (WallStreetBets – Kaggle)

3.2.1 Présentation de la source Reddit

Reddit est une plateforme de discussion en ligne organisée en communautés thématiques appelées *subreddits*. Le subreddit **WallStreetBets** est particulièrement connu pour ses discussions intenses autour des marchés financiers et des actions fortement spéculatives. Il regroupe des millions d'utilisateurs partageant analyses, opinions et stratégies d'investissement.

Dans le cadre de ce projet, les données Reddit ont été extraites à partir du dataset "**Reddit WallStreetBets Posts**" disponible sur la plateforme Kaggle. Ce dataset constitue une source fiable et structurée de données historiques issues des discussions de WallStreetBets.

3.2.2 Description du dataset Reddit

Le dataset Reddit utilisé est principalement composé de publications (*posts*) contenant des informations textuelles et temporelles. Il inclut notamment les champs suivants :

- identifiant du post,
- date et heure de publication,
- titre et contenu textuel,
- score du post (votes),
- nombre de commentaires.

Ces données textuelles représentent une matière première essentielle pour l'analyse de sentiment, permettant d'évaluer l'opinion générale des investisseurs à propos d'actions spécifiques.

3.2.3 Volume et période des données Reddit

Le volume des données Reddit est relativement important, comprenant plusieurs centaines de milliers de publications. La période couverte s'étend principalement de **2020 à 2021**, une période marquée par une forte volatilité des marchés et des événements majeurs tels que l'affaire GameStop.

Cette période est particulièrement pertinente pour étudier l'impact du sentiment social sur les marchés financiers.

3.2.4 Problèmes et défis liés aux données Reddit

Les données issues des réseaux sociaux présentent plusieurs défis :

- **Bruit textuel** : présence de messages non pertinents, d'abréviations, d'emojis ou de langage familier ;
- **Ambiguïté du langage** : ironie, sarcasme et exagération fréquents sur WallStreetBets ;
- **Déséquilibre des sentiments** : dominance de certains types d'opinions à certaines périodes ;
- **Temporalité** : décalage possible entre l'expression d'un sentiment et son impact réel sur le marché.

Ces défis nécessitent un prétraitement rigoureux avant toute analyse.

3.3 Données boursières (Stooq)

3.3.1 Présentation de la source Stooq

Les données financières utilisées dans ce projet proviennent de **Stooq**, une plateforme fournissant gratuitement des données boursières historiques pour un large éventail d'actions et de marchés. Stooq est largement utilisée dans les projets académiques en raison de la fiabilité et de la clarté de ses données.

3.3.2 Description des données boursières

Les données boursières collectées incluent principalement :

- le prix d'ouverture (*Open*),
- le prix le plus haut (*High*),
- le prix le plus bas (*Low*),

- le prix de clôture (*Close*),
- le volume de transactions (*Volume*).

Ces indicateurs, communément appelés **OHLCV**, constituent la base de nombreuses analyses financières et permettent de modéliser les variations de prix sur les marchés.

3.3.3 Volume et période des données boursières

Les données boursières couvrent la même période que les données Reddit afin de garantir une **cohérence temporelle**. Cette synchronisation est essentielle pour permettre la fusion des deux sources et l'analyse conjointe du sentiment social et des mouvements de marché.

Le volume de données est important mais reste structuré, facilitant leur exploitation dans un environnement Big Data.

3.3.4 Problèmes liés aux données boursières

Bien que structurées, les données boursières présentent également certaines contraintes :

- **Volatilité élevée** des prix, rendant la prédiction difficile ;
- **Présence de valeurs aberrantes**, notamment lors d'événements exceptionnels ;
- **Dépendance temporelle forte**, nécessitant des modèles capables de capturer les dynamiques séquentielles.

Ces caractéristiques justifient l'utilisation de modèles avancés de séries temporelles.

3.4 Fusion et cohérence des sources de données

Afin d'exploiter efficacement les deux sources de données, une étape de **synchronisation temporelle** est réalisée. Les données Reddit sont agrégées par intervalle de temps (par exemple par jour), tandis que les données boursières sont alignées sur la même granularité.

Cette fusion permet de créer un **dataset unifié**, combinant :

- les indicateurs financiers,
- les scores de sentiment issus de Reddit,
- les mesures d'activité sociale (nombre de posts, engagement).

Ce dataset constitue l'entrée principale des modèles de Machine Learning utilisés dans le projet.

3.5 Justification du choix des sources de données

Le choix de Reddit et de Stooq repose sur plusieurs critères :

- **Pertinence** : WallStreetBets est reconnu pour son influence sur les marchés financiers ;
- **Accessibilité** : les données sont disponibles publiquement et gratuitement ;
- **Complémentarité** : combinaison de données sociales et financières ;
- **Adaptation au Big Data** : volume et variété compatibles avec une architecture distribuée.

Ces sources offrent ainsi un cadre idéal pour expérimenter et valider une approche de prédiction boursière basée sur le sentiment social.

Chapitre 4 : Architecture du système Big Data

4.1 Vue d'ensemble de l'architecture

L'architecture du système proposé repose sur une **approche Big Data distribuée et modulaire**, conçue pour traiter des **flux de données hétérogènes en temps réel**. Elle vise à assurer la **scalabilité**, la **tolérance aux pannes** et la **séparation claire des responsabilités** entre les différents composants.

Le système est structuré en plusieurs couches fonctionnelles :

- **Couche de sources de données**
- **Couche de collecte**
- **Couche de streaming**
- **Couche de traitement**
- **Couche de stockage**
- **Couche Machine Learning**
- **Couche d'orchestration et de visualisation**

Cette architecture permet de traiter simultanément les données issues des réseaux sociaux (Reddit) et les données boursières (Stooq), puis de les exploiter à des fins de prédition.

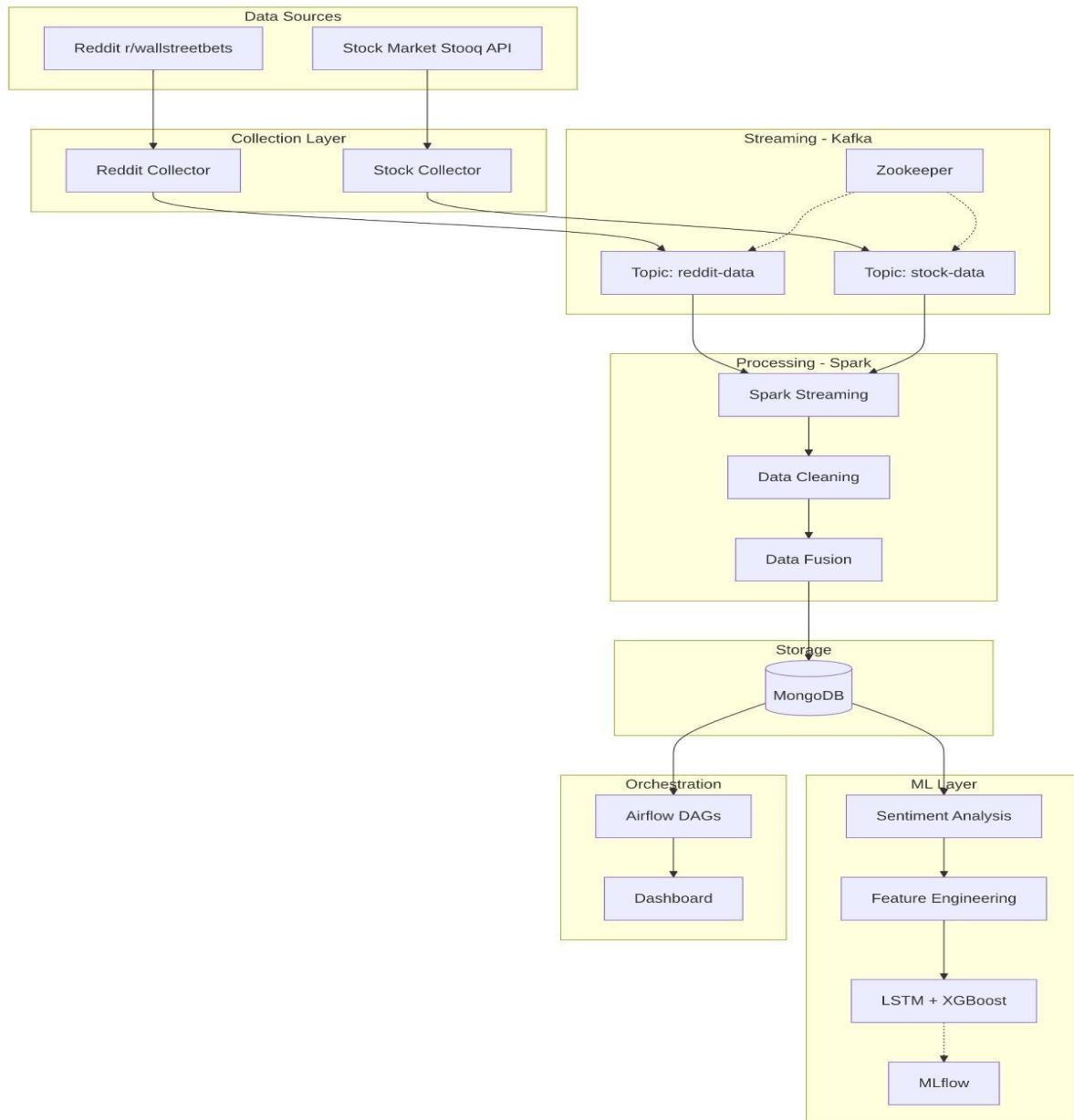


Figure 1 : Architecture globale du système Big Data

Cette figure présente l'architecture globale du système de prédiction boursière. Les données Reddit et boursières sont collectées, ingérées via Kafka, traitées par Spark Streaming, stockées dans MongoDB, puis exploitées par une couche Machine Learning et visualisées à travers un tableau de bord interactif.

4.2 Couche des sources de données

La couche des sources de données regroupe l'ensemble des fournisseurs d'information exploités par le système. Deux sources principales ont été retenues :

- **Reddit (WallStreetBets)** : fournit des données textuelles représentant le sentiment et l'activité des investisseurs particuliers.
- **Stooq** : fournit les données financières historiques et quasi temps réel des actions étudiées.

Ces sources sont complémentaires : les données Reddit apportent une dimension comportementale, tandis que les données Stooq reflètent la réalité du marché.

4.3 Couche de collecte des données

La couche de collecte est chargée de récupérer les données brutes depuis les sources externes et de les préparer pour l'ingestion dans le système de streaming.

Deux collecteurs distincts sont implémentés :

- un **collecteur Reddit**, chargé d'extraire les posts et métadonnées ;
- un **collecteur boursier**, chargé de récupérer les cours et volumes des actions.

Ces collecteurs sont développés en Python et assurent une première normalisation des données avant leur envoi vers Kafka.

4.4 Couche de streaming – Apache Kafka

Apache Kafka constitue le cœur de la couche de streaming. Il est utilisé comme **système de messagerie distribué** pour assurer une ingestion fiable et scalable des flux de données.

Deux topics principaux sont définis :

- reddit-data : pour les messages issus de Reddit ;
- stock-data : pour les données boursières.

Kafka permet de découpler les producteurs (collecteurs) des consommateurs (Spark), améliorant ainsi la robustesse globale du système.

La gestion du cluster Kafka est assurée par **Zookeeper**, qui coordonne les brokers et garantit la cohérence du système.

4.5 Couche de traitement – Apache Spark Streaming

La couche de traitement repose sur **Apache Spark Structured Streaming**, utilisé pour analyser les données en continu.

Les principales étapes du traitement sont :

- lecture des flux Kafka ;
- nettoyage et filtrage des données ;
- transformation des données textuelles Reddit ;
- agrégation temporelle ;
- fusion des données sociales et financières.

Spark permet un traitement distribué à faible latence, adapté aux contraintes du temps réel et aux volumes importants de données.

4.6 Couche de stockage – MongoDB

Les données traitées sont stockées dans **MongoDB**, une base de données NoSQL orientée documents. Ce choix est motivé par :

- la flexibilité du schéma ;
- la capacité à stocker des données hétérogènes ;
- la facilité d'intégration avec Spark et Python.

Plusieurs collections sont utilisées, notamment :

- données Reddit nettoyées ;
- données boursières enrichies ;
- données fusionnées destinées à l'apprentissage et à la visualisation.

4.7 Couche Machine Learning

La couche Machine Learning exploite les données stockées pour entraîner et évaluer les modèles de prédiction.

Elle comprend :

- une phase d'**analyse de sentiment** appliquée aux données Reddit ;
- une phase de **feature engineering** combinant indicateurs financiers et sociaux ;
- des modèles de **Deep Learning**, principalement des réseaux LSTM, ainsi que des modèles complémentaires (ex. XGBoost).

Les expériences et les modèles sont suivis et versionnés à l'aide de **MLflow**, garantissant la traçabilité des résultats.

4.8 Couche d'orchestration – Apache Airflow

Apache Airflow est utilisé pour orchestrer les différentes tâches du pipeline de données. Il permet de :

- planifier les traitements batch ;
- automatiser les étapes de préparation des données ;
- assurer le bon enchaînement des processus.

Les workflows sont définis sous forme de **DAGs**, facilitant la supervision et la maintenance du système.

4.9 Couche de visualisation

La couche de visualisation repose sur un **tableau de bord interactif développé avec Streamlit**. Elle permet :

- le suivi en temps réel des prix des actions ;
- l'analyse de l'activité Reddit ;
- la visualisation des prédictions et des performances du modèle.

Cette couche constitue l'interface finale entre le système et l'utilisateur.

4.10 Conteneurisation et déploiement – Docker

L'ensemble des composants du système est conteneurisé à l'aide de **Docker**. L'orchestration des services est assurée par **Docker Compose**, garantissant :

- un déploiement reproduitible ;
- une isolation des dépendances ;
- une mise en place rapide de l'environnement.

Cette approche facilite également le déploiement futur sur des infrastructures cloud.

```

Jan 19 6:32 PM
tmux
3 ③ generate_predictions.py ① 3 x  app.py ④ 2 x | ! docker-compose.yml x
Explorer 42/42
ml_models
  01_train_baseline_model.ipynb
  02_train_baseline_model.ipynb
  03_train_LSTM_Model.ipynb
  Untitled.ipynb
  __init__.py
  create_simple_model.py
  generate_predictions.py
orchestration
  airflow_dags
  __init__.py
  app.py
  mongo.py
  predictor_service.py
  predict.py
  relayer_simulator.py
volumes
README.md
$ activate_env.sh
$ debug_csv.py
! docker-compose.yml
$ init-kafka.sh
$ requirements.txt
$ start_relayer.sh
NORMAL > feature/integration > docker-compose.yml
[0] 1:bash 2:bash 3:nvim* 4:python3.11 5:docker- 6:bash

```

The screenshot shows a tmux session with multiple panes. The left pane displays a file browser with a tree view of files and folders, including 'ml_models' and 'orchestration' directories containing various Python scripts and notebooks. The right pane shows a terminal window with the command 'docker-compose.yml' highlighted. The terminal content is a Docker Compose configuration file:

```

services:
  # Zookeeper - Required for Kafka
  zookeeper:
    image: confluentinc/cp-zookeeper:7.5.0
    hostname: zookeeper
    container_name: zookeeper
    ports:
      - "2181:2181"
    environment:
      ZOOKEEPER_CLIENT_PORT: 2181
      ZOOKEEPER_TICK_TIME: 2000
    networks:
      - bigdata-network
    healthcheck:
      test: ["CMD", "nc", "-z", "localhost", "2181"]
      interval: 10s
      timeout: 5s
      retries: 5

  # Kafka - Message Streaming
  kafka:
    image: confluentinc/cp-kafka:7.5.0
    hostname: kafka
    container_name: kafka
    depends_on:
      - zookeeper
      | condition: service_healthy

```

Figure 2 : Orchestration des services Big Data avec Docker Compose

Cette figure illustre un extrait du fichier de configuration Docker Compose utilisé pour orchestrer les différents services de l'architecture Big Data. Elle montre notamment la configuration de Zookeeper et Kafka, essentiels à la gestion du streaming de données. Cette approche permet un déploiement reproduitible, modulaire et facilement extensible du système.

Tableau 1 : Rôle des composants de l'architecture

Composant	Rôle principal
Reddit / Stooq	Sources de données
Collecteurs Python	Extraction des données
Kafka	Ingestion et streaming
Spark Streaming	Traitemet temps réel
MongoDB	Stockage des données
XGBOOST/ML	Prédiction boursière
Airflow	Orchestration
Streamlit	Visualisation

Conclusion du chapitre

Ce chapitre a présenté l'architecture Big Data mise en place pour répondre aux exigences de traitement en temps réel, de scalabilité et de robustesse. Cette architecture constitue la colonne vertébrale du système et permet l'intégration fluide des données sociales et financières. Le chapitre suivant détaille le **pipeline de traitement des données**, depuis l'ingestion jusqu'à la préparation des données pour la modélisation.

Chapitre 5 : Pipeline de traitement des données

5.1 Objectifs du pipeline de traitement

Le pipeline de traitement des données constitue un élément central du système de prédiction boursière proposé. Son objectif principal est de **transformer des données brutes hétérogènes** — issues à la fois des marchés financiers et des réseaux sociaux — en **données propres, structurées et exploitables** par les modèles de Machine Learning.

Dans un contexte Big Data et temps réel, ce pipeline doit répondre à plusieurs contraintes :

- gérer des flux continus de données à forte vitesse ;
- assurer une qualité élevée des données ;
- permettre la fusion de sources de nature différente ;
- garantir la scalabilité et la robustesse du traitement.

5.2 Ingestion des données en temps réel

5.2.1 Ingestion des données Reddit

Les données Reddit sont ingérées sous forme de flux via **Apache Kafka**. Les publications issues de WallStreetBets sont envoyées dans le topic reddit-data par un producteur Kafka développé en Python.

Chaque message contient notamment :

- l'identifiant du post,
- le contenu textuel,
- la date et l'heure de publication,
- des métadonnées telles que le score ou le nombre de commentaires.

Cette approche permet une ingestion continue et asynchrone, essentielle pour l'analyse du sentiment en temps réel.

5.2.2 Ingestion des données boursières

Les données boursières sont collectées à partir de la source Stooq et transmises vers Kafka dans le topic stock-data.

Chaque message contient les informations OHLCV associées à une action donnée, accompagnées d'un horodatage précis.

L'utilisation de Kafka permet de synchroniser les flux financiers avec les flux sociaux, facilitant ainsi leur traitement conjoint.

5.3 Nettoyage et prétraitement des données

5.3.1 Nettoyage des données Reddit

Les données textuelles issues de Reddit nécessitent un nettoyage approfondi avant toute analyse. Les principales opérations réalisées sont :

- suppression des URLs, caractères spéciaux et emojis ;
- conversion du texte en minuscules ; • suppression des mots vides (*stop words*) ;
- normalisation du texte.

Ces étapes permettent de réduire le bruit et d'améliorer la qualité de l'analyse de sentiment.

5.3.2 Nettoyage des données boursières

Les données boursières sont généralement structurées, mais nécessitent également un prétraitement, notamment :

- gestion des valeurs manquantes ;
- suppression des doublons ;
- détection et traitement des valeurs aberrantes.

Ces opérations assurent la cohérence des séries temporelles utilisées par les modèles.

5.4 Analyse de sentiment des posts Reddit

Une analyse de sentiment est appliquée aux textes Reddit à l'aide de **VADER (Valence Aware Dictionary and sEntiment Reasoner)**. VADER attribue un score **compound** compris entre -1 et +1, ainsi que des scores **positif / négatif / neutre**. Ces scores sont ensuite agrégés par intervalle de temps (ex. par jour) pour représenter le sentiment collectif des investisseurs à un instant donné, et sont fusionnés avec les données boursières pour la phase de modélisation.

5.5 Agrégation temporelle des données

Pour permettre la fusion des données sociales et financières, une étape d'**agrégation temporelle** est nécessaire.

Les données Reddit, initialement très fines (à la minute ou à l'heure), sont agrégées selon la même granularité que les données boursières.

Les indicateurs agrégés incluent :

- score moyen de sentiment ;
- nombre de publications ;
- niveau moyen d'engagement.

Cette agrégation facilite la comparaison et la fusion des deux sources.

5.6 Fusion des données sociales et financières

La fusion des données constitue une étape clé du pipeline. Elle consiste à combiner : •

- les données boursières (prix, volume, indicateurs techniques) ;
- les données issues de l'analyse de sentiment Reddit.

Cette fusion est réalisée sur la base de l'horodatage, garantissant une cohérence temporelle entre les deux sources.

Le résultat est un **dataset unifié**, représentant à la fois l'état du marché et le sentiment des investisseurs.

5.7 Feature engineering

À partir des données fusionnées, une phase de **feature engineering** est réalisée afin de créer des variables pertinentes pour l'apprentissage des modèles.

Les principales features incluent :

- rendements journaliers ; • moyennes mobiles ;
- indicateurs de volatilité ;
- scores de sentiment ;
- mesures d'activité sociale.

Ces features permettent aux modèles de capturer des relations complexes entre les données financières et sociales.

5.8 Stockage des données traitées

Les données finales issues du pipeline sont stockées dans **MongoDB**.

Plusieurs collections sont définies afin de séparer :

- les données brutes ;
 - les données nettoyées ;
 - les données enrichies et fusionnées.

Cette organisation facilite l'accès aux données pour les phases d'entraînement, de prédiction et de visualisation.

5.9 Gestion des erreurs et robustesse du pipeline

Dans un environnement temps réel, la gestion des erreurs est essentielle. Le pipeline intègre plusieurs mécanismes pour garantir sa robustesse :

- checkpoints Spark pour éviter la perte de données ;
 - gestion des exceptions lors du traitement ;
 - journalisation des erreurs pour faciliter le débogage.

Ces mécanismes assurent la continuité du traitement même en cas de défaillance partielle.

5.10 Résumé du pipeline de traitement

Le pipeline de traitement permet de transformer des flux de données brutes en informations structurées et exploitables par les modèles de Machine Learning. Il joue un rôle fondamental dans la performance globale du système et constitue le lien entre l'architecture Big Data et la phase de modélisation.



```
Jan 19 03:17PM
```

```
producer_tutorial.ipynb ① reddit_pipeline.py ② stock_pipeline.py ③ 61 X
```

```
producer_tutorial.ipynb ① reddit_pipeline.py ② stock_pipeline.py ③ 61 X
```

```
from pyspark import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
from datetime import datetime, timedelta
import numpy as np
import pandas as pd
# pandas imported but unused
Pyright: "pd" is not a module
```

```
# CONFIG
```

```
KAFKA_BOOTSTRAP_SERVER = "kafka:kafka:9092"
MONGO_URI = "mongodbs://mongodbs:27017"
MONGO_DB = "stockmarket_db"
```

```
STOCK_RAW_CHECKPOINT = "/tmp/chk/stck_raw"
STOCK_FEATURE_CHECKPOINT = "/tmp/chk/stck_stock_features"
```

```
# SPARK SESSION
```

```
spark = (
    SparkSession.builder.appName("StockContinuousPipeline")
    .config(
        "spark.jars.packages",
        "org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.0",
        "org.mongodb.spark:mongo-spark-connector_2.12:10.3.0",
    )
)
```

```
NORMAL ④ Feature/integration ⑤ 61 ⑥ data_processing/stock_pipeline.py ⑦ gg ⑧ 6 ⑨ 437 ⑩ 21 ⑪ 183 ⑫ 10 ⑬ 136 ⑭ 181 ⑮ 183
```

```
(1B) lib@192-168-1-10:~/dev/reddit/ ① reddit_pipeline.py ② stock_pipeline.py ③ 61 X
```

Figure 3 : Pipeline de traitement des données

Cette figure illustre le pipeline de traitement des données boursières implémenté à l'aide d'Apache Spark Structured Streaming. Les données sont consommées depuis Apache Kafka, nettoyées et transformées en continu, puis stockées dans MongoDB. Des mécanismes de checkpointing sont utilisés afin d'assurer la tolérance aux pannes et la reprise automatique du traitement en cas d'erreur.

```

Jan 19 6:30 PM      tmux
[1] 28/28 * reddit_pipeline.py & (2) *
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
import re
# -- CONFIG --
# Known Tickers
KNOWN_TICKERS = ["GOOG", "AMC", "TSLA", "AAPL", "BBB", "NOK", "PLTR", "SPCE"]
KNOWN_TICKERS_SET = set(KNOWN_TICKERS)
# Kafka Bootstrap Server
KAFKA_BOOTSTRAP_SERVER = "kafka:9992"
MONGO_URI = "mongodb://mongod:27017"
# Checkpoints
RAW_CHECKPOINT = "/tmp/chk/reddit_raw"
FEATURE_CHECKPOINT = "/tmp/chk/reddit_features"
# SPARK SESSION - FIXED: Added required packages
spark = SparkSession.builder.appName("RedditContinuousPipeline") \
    .config("spark.jars.packages", \
        "org.apache.spark:spark-sql-kafka-0-10_2:1.3.3,5.0," \
        "org.mongodb.spark:mongo-spark-connector_2.12:10.1.0")

```

Figure 4 : Pipeline Spark pour le traitement des données Reddit

Cette figure montre le pipeline Spark utilisé pour traiter les données boursières. Les données sont nettoyées, enrichies par des indicateurs financiers et stockées sous une forme exploitable par les modèles de Machine Learning.

Conclusion du chapitre

Ce chapitre a détaillé l'ensemble des étapes du pipeline de traitement des données, depuis l'ingestion en temps réel jusqu'à la création des features. Le chapitre suivant se concentre sur la **modélisation et le Machine Learning**, en présentant les modèles utilisés pour la prédiction boursière.

Chapitre 6 : Modélisation et Machine Learning

6.1 Objectifs de la modélisation

L'objectif principal de la phase de modélisation est de **prédirer l'évolution des cours boursiers** à partir d'un ensemble de variables combinant des **données financières historiques** et des **indicateurs de sentiment issus des réseaux sociaux**. Cette phase vise à :

- exploiter la dépendance temporelle des séries financières ;
- intégrer l'impact comportemental des investisseurs ;
- produire des prédictions fiables à court terme ;
- évaluer rigoureusement les performances des modèles.

6.2 Choix des modèles de prédiction

Le modèle retenu est **XGBoost**, un algorithme de **Gradient Boosting** performant pour les tâches de régression et de classification, particulièrement adapté aux données tabulaires enrichies par des features financières et de sentiment.

6.3 Présentation du modèle XGBOOST

Le modèle **XGBoost (eXtreme Gradient Boosting)** est un algorithme de **Machine Learning basé sur le principe du gradient boosting**, qui combine un ensemble d'arbres de décision afin d'améliorer progressivement les performances de prédiction. Chaque nouvel arbre est entraîné pour corriger les erreurs commises par les modèles précédents, ce qui permet d'obtenir un modèle robuste et précis. XGBoost est particulièrement adapté aux **données tabulaires** et hétérogènes, telles que les données financières enrichies par des **indicateurs de sentiment**, car il est capable de capturer des **relations non linéaires complexes** entre les variables. De plus, il intègre des mécanismes de régularisation permettant de limiter le surapprentissage et d'améliorer la généralisation du modèle.

Grâce à ces caractéristiques, XGBoost permet de **modéliser efficacement les tendances, les variations et les ruptures** présentes dans les séries financières, tout en exploitant l'impact du sentiment issu des réseaux sociaux.

6.4 Données d'entrée du modèle

Les données utilisées pour l'entraînement du modèle sont issues du **dataset fusionné** décrit dans le chapitre précédent.

Elles comprennent notamment :

Variables financières

- prix de clôture,
- volume,
- rendements journaliers,
- moyennes mobiles.

Variables sociales

- score moyen de sentiment Reddit,
- nombre de posts par période,
- niveau d'engagement.

Ces variables sont normalisées afin d'améliorer la convergence du modèle.

6.5 Préparation des données pour l'apprentissage

Avant l'entraînement, plusieurs étapes sont nécessaires :

6.5.1 Création des séquences temporelles

Les données sont découpées en **fenêtres temporelles glissantes**, où chaque séquence contient un nombre fixe de pas de temps (par exemple 30 jours).

Chaque séquence est associée à une valeur cible correspondant au prix futur de l'action.

6.5.2 Séparation des jeux de données

Les données sont divisées en :

- jeu d'entraînement,
- jeu de validation,
- jeu de test.

Cette séparation respecte l'ordre temporel afin d'éviter toute fuite d'information.

6.6 Architecture du modèle XGBOOT

Le modèle de prédiction implémenté repose sur **XGBoost (eXtreme Gradient Boosting)**, un algorithme de Machine Learning basé sur l'agrégation séquentielle de **plusieurs arbres de décision**. Le principe consiste à entraîner chaque nouvel arbre afin de corriger les erreurs des arbres précédents, ce qui permet d'améliorer progressivement la précision globale du modèle.

L'architecture du modèle XGBoost utilisée dans ce projet comprend :

- un **ensemble d'arbres de décision** construits de manière itérative,
- un mécanisme de **gradient boosting** pour l'optimisation,
- des techniques de **régularisation** (L1 et L2) afin de limiter le surapprentissage,
- une fonction objectif adaptée aux tâches de **régression boursière**.

L'optimisation du modèle est réalisée à partir de la **fonction de perte quadratique moyenne (MSE)**, tandis que les hyperparamètres tels que le nombre d'arbres, la profondeur maximale et le taux d'apprentissage sont ajustés afin d'obtenir un bon compromis entre performance et généralisation.

Cette architecture permet au modèle de **capturer efficacement les relations non linéaires**, les tendances et les variations présentes dans les données financières enrichies par les **indicateurs de sentiment VADER**.

6.7 Entraînement du modèle

L'entraînement est réalisé sur plusieurs époques afin de permettre au modèle de converger vers une solution optimale.

Des mécanismes tels que l'**early stopping** sont utilisés pour éviter le surapprentissage.

L'évolution de la perte est suivie tout au long de l'entraînement afin d'analyser la stabilité et la performance du modèle.

6.8 Modèles complémentaires et comparaison

En complément du modèle principal **XGBoost**, des approches plus simples ont été testées afin de servir de **références de comparaison**. Parmi celles-ci figurent :

- des **modèles baselines simples**, tels que la moyenne mobile,
- des méthodes de prédiction basées uniquement sur les **données financières historiques**, sans intégration du sentiment.

Cette comparaison permet de **mettre en évidence l'apport de XGBoost**, notamment sa capacité à exploiter efficacement des **features agrégées** issues à la fois des données boursières et des **indicateurs de sentiment VADER**, et à capturer des relations non linéaires complexes influençant les mouvements du marché.

6.9 Métriques d'évaluation

Les performances des modèles sont évaluées à l'aide de plusieurs métriques :

- **RMSE (Root Mean Squared Error)** : mesure l'erreur moyenne de prédiction ;
- **MAE (Mean Absolute Error)** : évalue l'erreur absolue ;
- **Accuracy directionnelle** : capacité à prédire correctement la hausse ou la baisse du prix.

Ces métriques permettent une évaluation complète, à la fois quantitative et qualitative.

6.10 Suivi des expériences avec MLflow

Afin d'assurer la traçabilité des expériences, le projet intègre **MLflow** pour :

- enregistrer les hyperparamètres ;
- suivre les métriques d'évaluation ;
- versionner les modèles entraînés.

Cette approche facilite la comparaison des différentes configurations et améliore la reproductibilité du projet.

6.11 Résumé de la phase de modélisation

La phase de modélisation constitue le cœur intelligent du système. L'utilisation de **XGBoost** permet d'exploiter efficacement les variables financières et les indicateurs de sentiment, tout en offrant de bonnes performances et une interprétabilité partielle via l'importance des features.

Conclusion du chapitre

Ce chapitre a présenté les choix méthodologiques et techniques relatifs à la modélisation et au Machine Learning. Le chapitre suivant se concentre sur **l'implémentation technique**, en détaillant l'organisation du code et l'intégration des différents composants du système.

Chapitre 7 – Implémentation technique

7.1 Introduction

Ce chapitre présente l'implémentation technique du système de prédiction boursière en temps réel. Il décrit brièvement les technologies utilisées et la manière dont les différents composants ont été intégrés afin de mettre en œuvre l'architecture Big Data présentée précédemment.

7.2 Choix technologiques

Le système a été développé en s'appuyant sur des technologies Big Data modernes et largement utilisées dans l'industrie.



Python est utilisé comme socle principal pour la collecte des données, le traitement et l'implémentation des modèles de Machine Learning.
L'ingestion des données en temps réel est assurée par

Figure 5 : Python logo



Apache Kafka, qui permet de gérer efficacement les flux continus issus des données boursières et des discussions Reddit.
Le traitement distribué des données est réalisé à l'aide de

Figure 6 : Apache kafka logo



Apache Spark Structured Streaming, offrant de bonnes performances et une grande scalabilité. Les données traitées sont ensuite stockées dans

Figure 7 : Apach Spark Structured Streaming



MongoDB, une base de données NoSQL adaptée aux données hétérogènes et semi-structurées.

Figure 8 : MongoDB .BD NosSQL

7.3 Déploiement et orchestration

L'ensemble des composants du système est conteneurisé à l'aide de



Docker, garantissant une isolation correcte des services et une meilleure reproductibilité de l'environnement.

Figure 9 : docker logo

L'orchestration est réalisée via



Docker Compose, facilitant le lancement et la gestion des différents services tels que Kafka, Spark, MongoDB et les modules applicatifs.

Figure 10 : docker compose

Cette approche permet un déploiement rapide et cohérent du système, aussi bien en environnement local que dans un contexte de démonstration.

7.4 Intégration des composants

Les différents modules du système communiquent de manière fluide :

- les producteurs Kafka envoient les données vers les topics dédiés,
- Spark consomme ces flux, applique les traitements nécessaires et stocke les résultats,
- les données finales sont rendues accessibles aux autres composants du système.

Cette intégration assure un flux de données continu depuis la collecte jusqu'à l'exploitation finale.

Conclusion du chapitre

Ce chapitre a présenté l'implémentation technique du système de prédiction boursière. Il montre comment l'architecture Big Data a été concrètement mise en œuvre à travers des outils adaptés, permettant un traitement en temps réel, une bonne scalabilité et une intégration efficace des différents composants.

Chapitre 8 : Visualisation et tableau de bord

8.1 Objectifs de la visualisation

La visualisation des données constitue une étape essentielle pour faciliter l'interprétation des résultats du système de prédiction boursière. L'objectif principal du tableau de bord est de fournir une **vue synthétique, claire et interactive** de l'évolution des marchés financiers, du sentiment exprimé sur Reddit et des prédictions générées par les modèles de Machine Learning.

Le tableau de bord permet ainsi aux utilisateurs d'analyser simultanément les données financières réelles, l'activité sociale et les résultats prédictifs.

8.2 Outil de visualisation

Le tableau de bord a été développé à l'aide de **Streamlit**, un **framework** Python dédié à la création d'interfaces interactives pour les applications de Data Science. **Streamlit** permet un affichage dynamique des données et une mise à jour quasi temps réel des indicateurs, tout en restant simple à intégrer avec **MongoDB** et les pipelines existants.

8.3 Indicateurs clés affichés (KPI)

Le dashboard affiche plusieurs indicateurs clés de performance, notamment :

- le prix actuel de l'action étudiée,
- la variation journalière du prix,
- le volume de transactions,
- le score moyen de sentiment Reddit,
- la précision directionnelle du modèle de prédiction.

Ces indicateurs offrent une vision rapide de l'état du marché et de la performance du modèle.

8.4 Visualisation des données boursières

Les données boursières sont représentées sous forme de graphiques temporels permettant d'analyser l'évolution des prix et des volumes.

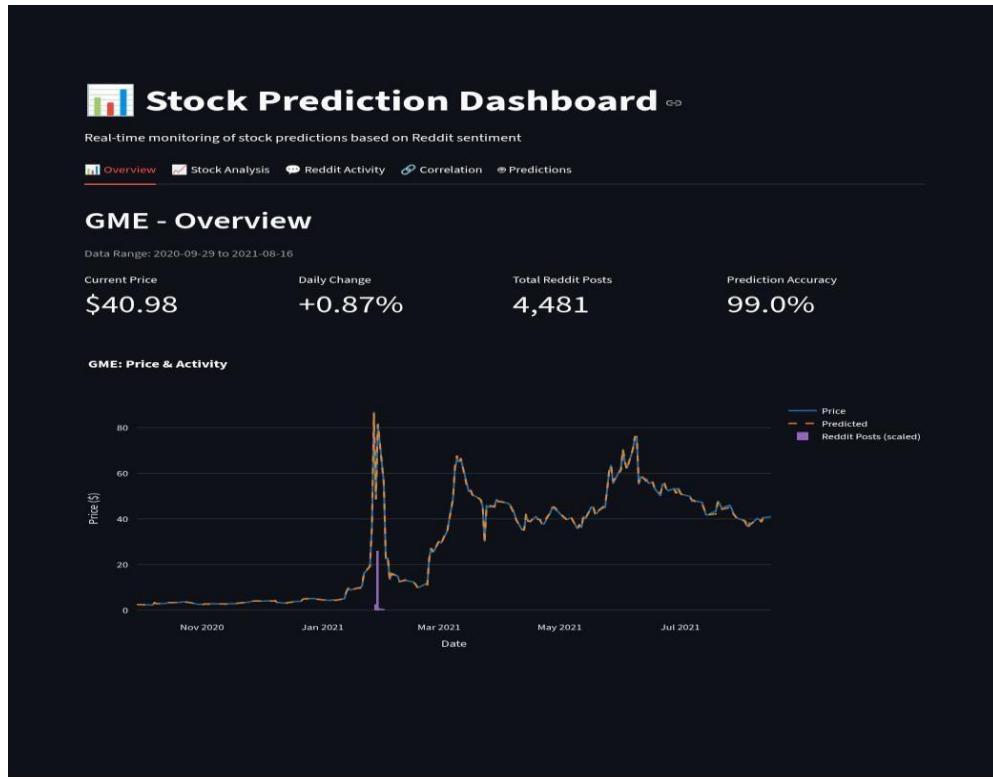


Figure 11 : Vue d'ensemble du tableau de bord

Cette figure présente la vue globale du tableau de bord interactif. Elle regroupe les principaux KPI ainsi que les graphiques de prix et d'activité, offrant une vision synthétique du marché et des prédictions.



Figure 12 : Évolution du prix de l'action

Cette figure illustre l'évolution du prix de l'action étudiée sur la période considérée. Elle permet d'identifier les tendances haussières et baissières ainsi que les périodes de forte volatilité.

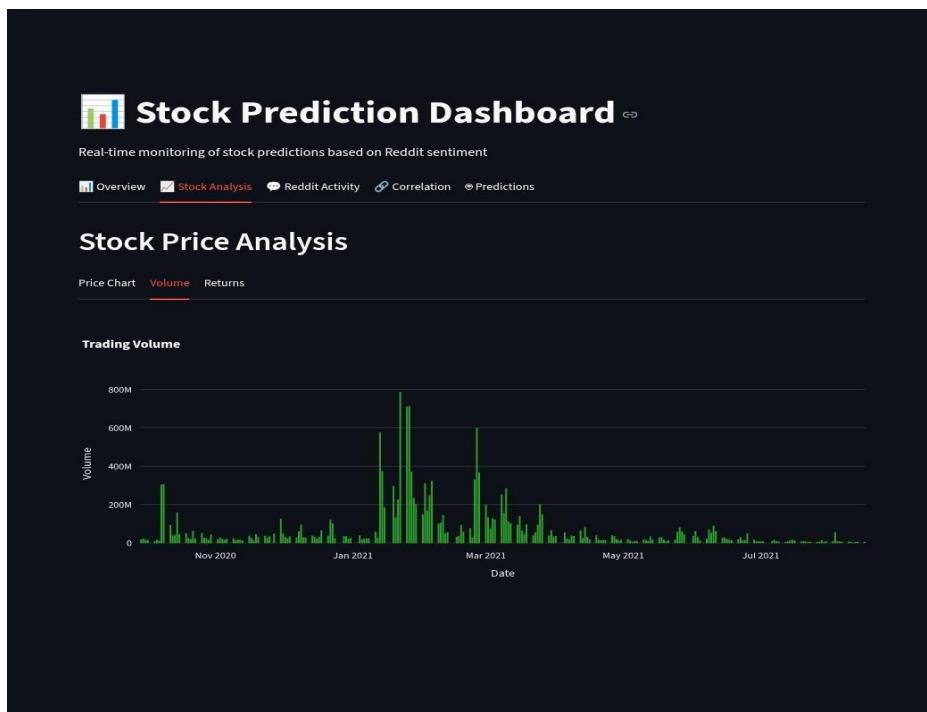


Figure 13 : Analyse du volume de transactions

Cette figure représente le volume de transactions associé à l'action. Les pics de volume mettent en évidence des phases d'activité intense sur le marché.

8.5 Visualisation de l'activité et du sentiment Reddit

Le tableau de bord intègre également des visualisations dédiées aux données sociales afin de mesurer l'impact du sentiment des investisseurs.

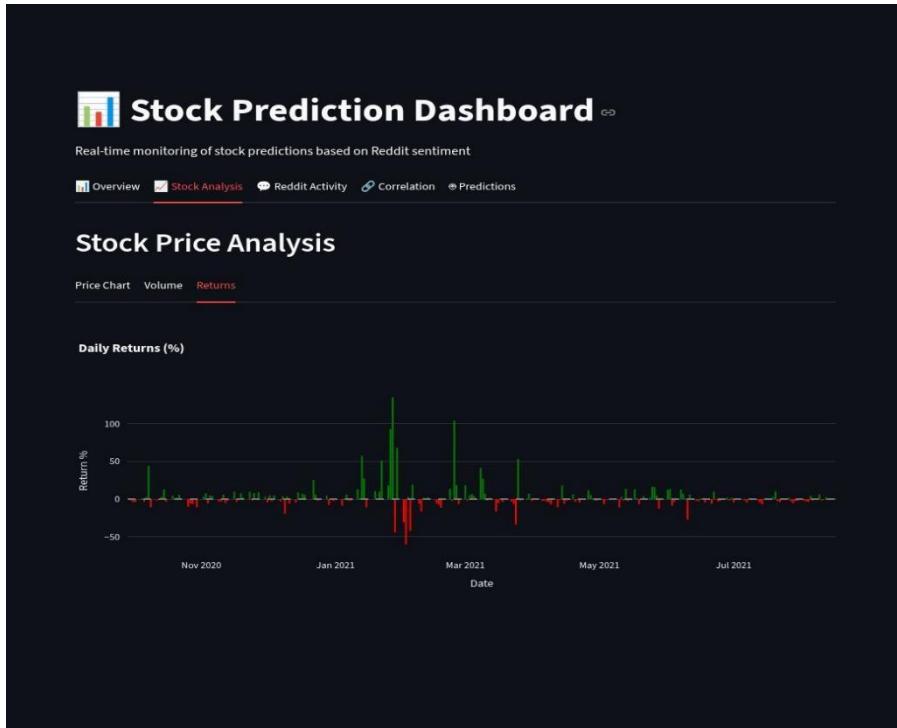


Figure 14 : Fréquence des publications Reddit

Cette figure montre l'évolution du nombre de publications Reddit liées à l'action étudiée. Les pics observés correspondent à des périodes d'intérêt accru des investisseurs particuliers.

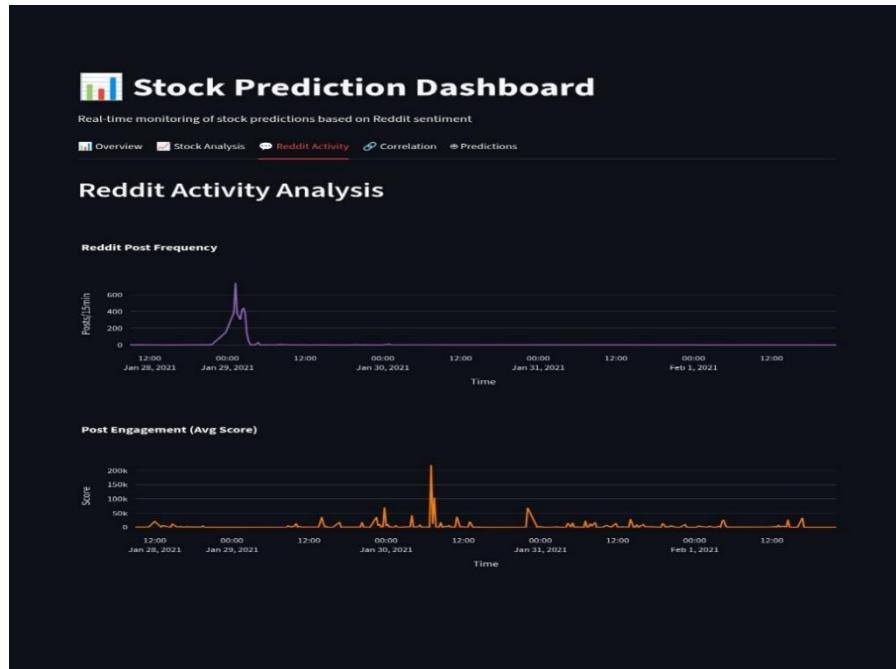


Figure 15 : Score de sentiment Reddit

Cette figure illustre l'évolution du score moyen de sentiment extrait des discussions Reddit. Elle permet d'analyser la relation entre l'opinion collective et les mouvements du marché.

8.6 Visualisation des prédictions

Les résultats du modèle de Machine Learning sont intégrés directement dans le tableau de bord afin de comparer les prédictions aux valeurs réelles.

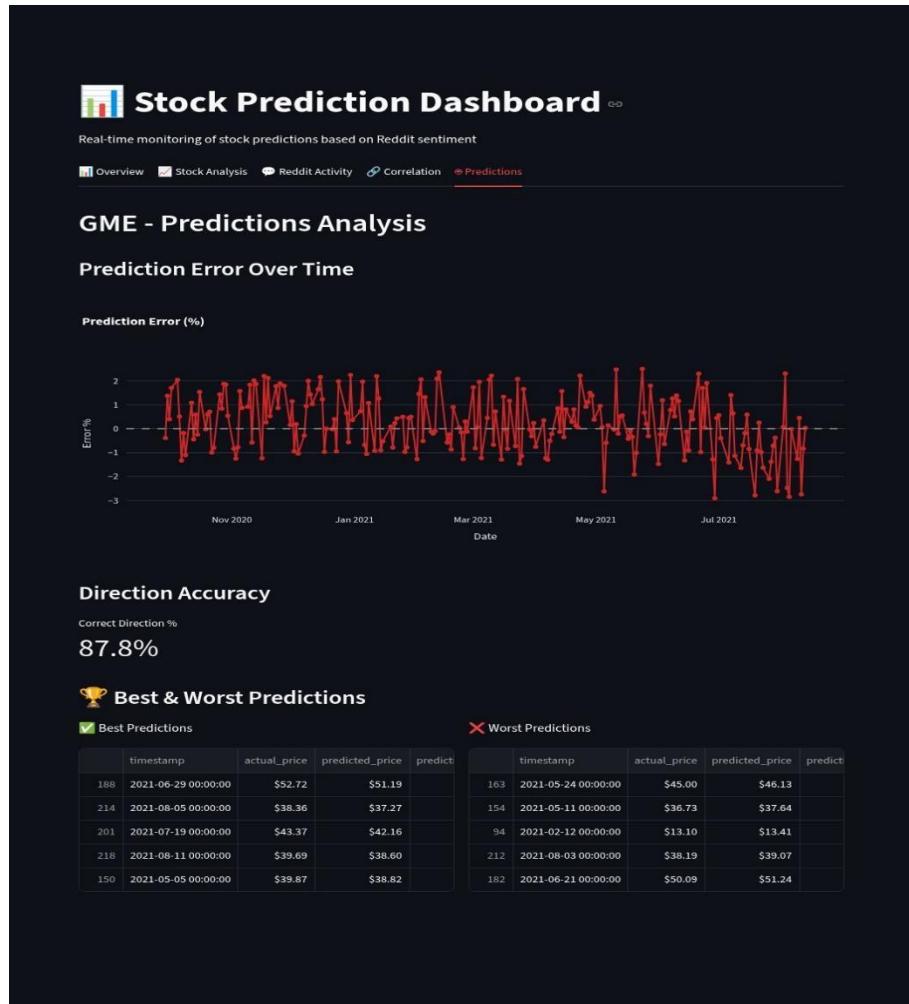


Figure 16 : Comparaison prix réel vs prix prédit & Erreur de prédition dans le temps

Cette figure présente la comparaison entre le prix réel de l'action et le prix prédict par le modèle LSTM. Elle met en évidence la capacité du modèle à suivre la dynamique générale du marché.

Montre l'évolution de l'erreur de prédition au fil du temps, permettant d'évaluer la stabilité et la fiabilité du modèle.

8.7 Conclusion du chapitre

Ce chapitre a présenté le tableau de bord interactif développé pour la visualisation des données financières, du sentiment Reddit et des résultats de prédition. Grâce à des graphiques clairs et des indicateurs synthétiques, le dashboard constitue un outil efficace pour l'analyse et la prise de décision, aussi bien dans un cadre académique que démonstratif.

Chapitre 9 : Résultats et analyse

9.1 Résultats des prédictions

Le modèle **XGBoost** a montré de bonnes performances prédictives en exploitant conjointement les **données boursières** et les **scores de sentiment VADER**, confirmant la pertinence de l'approche proposée.

9.2 Analyse de l'impact du sentiment Reddit

L'intégration des indicateurs de sentiment issus de Reddit a permis d'enrichir significativement les données d'entrée du modèle. Les analyses montrent que les **pics d'activité et de sentiment** sur WallStreetBets coïncident souvent avec des mouvements importants des prix des actions.

Cette observation confirme que le sentiment des investisseurs particuliers constitue un **facteur pertinent** pour l'anticipation des variations boursières à court terme, notamment pour des actions fortement médiatisées.

9.3 Étude de cas (GME / AMC / TSLA)

L'étude de cas réalisée sur des actions telles que **GameStop (GME)**, **AMC** et **Tesla (TSLA)** met en évidence une relation notable entre l'activité sociale et la dynamique du marché.

Dans le cas de GME, par exemple, les périodes de forte discussion sur Reddit correspondent à des fluctuations marquées des prix, que le modèle parvient partiellement à anticiper.

9.4 Discussion des performances

Les résultats obtenus indiquent que le modèle présente une **bonne précision directionnelle**, ce qui est particulièrement intéressant dans un contexte d'aide à la décision financière. Cependant, la prédiction exacte des valeurs reste limitée par la complexité intrinsèque des marchés et par l'influence d'événements exogènes non observables dans les données.

Conclusion du chapitre

Ce chapitre a présenté et analysé les résultats du système de prédiction boursière. Les expériences menées montrent que la combinaison des données financières et du sentiment Reddit, associée à des modèles de Deep Learning, permet d'obtenir des résultats prometteurs tout en soulignant les défis liés à la volatilité des marchés.

Chapitre 10 : Limites du projet

10.1 Limites techniques

Le système repose sur une architecture Big Data fonctionnelle, mais certaines limites techniques subsistent. Le traitement en temps réel dépend fortement des performances de l'infrastructure matérielle et peut être affecté en cas de forte montée en charge. De plus, l'exécution locale via Docker limite la scalabilité par rapport à une infrastructure cloud distribuée.

10.2 Limites liées aux données

Les données issues de Reddit sont intrinsèquement bruitées et subjectives. Le langage utilisé sur WallStreetBets inclut souvent de l'ironie, de l'exagération ou du sarcasme, ce qui peut affecter la précision de l'analyse de sentiment.

Par ailleurs, certaines variations boursières sont causées par des événements externes (annonces économiques, décisions politiques) qui ne sont pas directement reflétés dans les données sociales.

10.3 Limites des modèles

Bien que le modèle **XGBoost** offre de bonnes performances pour la prédiction boursière, il reste sensible aux **fortes variations imprévisibles du marché**. La prédiction exacte des prix demeure complexe en raison de la volatilité et des événements exogènes, et le modèle se révèle généralement **plus performant pour anticiper la tendance générale des mouvements** que la valeur précise du prix à un instant donné.

Conclusion du chapitre

Ce chapitre a mis en évidence les principales limites du projet, tant sur le plan technique que méthodologique. Ces limites ouvrent la voie à plusieurs pistes d'amélioration présentées dans le chapitre suivant.

Chapitre 11 : Perspectives et améliorations

11.1 Enrichissement des sources de données

Une première amélioration consisterait à intégrer de nouvelles sources de données, telles que Twitter, les actualités financières ou les indicateurs macroéconomiques, afin de mieux capturer l'ensemble des facteurs influençant les marchés.

11.2 Amélioration des modèles de prédiction

L'utilisation de modèles plus avancés, tels que les **Transformers** ou les modèles à mécanisme d'attention, pourrait améliorer la capacité du système à capturer des relations complexes et à long terme dans les données.

11.3 Déploiement et passage à l'échelle

Le déploiement du système sur une infrastructure cloud (AWS, Azure ou GCP) permettrait d'améliorer la scalabilité, la disponibilité et la robustesse globale de la solution.

11.4 Fonctionnalités avancées

Des fonctionnalités supplémentaires pourraient être ajoutées, telles que :

- un système d'alertes en temps réel,
- des recommandations d'achat ou de vente,
- une personnalisation du tableau de bord selon le profil de l'utilisateur.

Conclusion du chapitre

Ce chapitre a présenté plusieurs perspectives d'évolution visant à renforcer la performance, la robustesse et la portée du système de prédiction boursière.

Conclusion générale

Ce projet a permis de concevoir et de mettre en œuvre un **système de prédition boursière en temps réel** combinant des données financières et le sentiment issu des réseaux sociaux, en particulier Reddit. En s'appuyant sur une architecture Big Data basée sur Kafka, Spark, MongoDB et Docker, le système assure une collecte, un traitement et une visualisation efficaces des données.

Bien que le modèle **XGBoost** offre de bonnes performances pour la prédition boursière, il reste sensible aux **fortes variations imprévisibles du marché**. La prédition exacte des prix demeure complexe en raison de la volatilité et des événements exogènes, et le modèle se révèle généralement **plus performant pour anticiper la tendance générale des mouvements** que la valeur précise du prix à un instant donné.

Ce travail met en évidence la pertinence de la combinaison **Big Data + Intelligence Artificielle** pour l'analyse des marchés financiers, tout en soulignant les défis liés à la volatilité et à la qualité des données. Il constitue ainsi une base solide pour des améliorations futures et des déploiements à plus grande échelle.

Références bibliographiques

Documentation officielle et supports pédagogiques

- **Supports de cours du professeur Big Data NonSQL**
<https://boti.education/u/ismagi/extranet/eleve/devoirs>
- **Apache Kafka – Concepts et architecture (cours et travaux dirigés)**
- **Apache Spark – Traitement distribué et Structured Streaming (supports de cours)**
- **MongoDB – Bases NoSQL et intégration Big Data (cours)**
- **TensorFlow & Deep Learning – Réseaux de neurones et LSTM (cours et TP) • Streamlit – Visualisation des données (support pédagogique)**

Sources de données

Kaggle – Reddit WallStreetBets Dataset

<https://www.kaggle.com/datasets/gpreda/reddit-wallstreetbets-posts>

Stooq – Historical Stock Market Data <https://stooq.com/>

Résumé global du rapport

Ce rapport présente la conception et la réalisation d'un **système de prédition boursière en temps réel** exploitant à la fois les **données financières** et le **sentiment des investisseurs exprimé sur les réseaux sociaux**. Le projet s'inscrit dans un contexte de **forte volatilité des marchés financiers**, où l'information et l'opinion collective jouent un rôle déterminant dans l'évolution des prix.

Après une étude des approches existantes, une **architecture Big Data** a été proposée et implémentée. Elle repose sur **Apache Kafka** pour l'ingestion des données en streaming, **Apache Spark Structured Streaming** pour le traitement distribué, **MongoDB** pour le stockage et **Docker** pour la conteneurisation et le déploiement. Les données Reddit et boursières sont nettoyées, fusionnées et transformées afin de produire des jeux de données exploitables par des modèles de **Machine Learning**.

La modélisation s'appuie sur **XGBoost**, un algorithme de **Gradient Boosting** particulièrement adapté aux données tabulaires enrichies par des **indicateurs de sentiment**. L'analyse de sentiment est réalisée à l'aide de **VADER**, permettant d'extraire des scores représentant l'opinion collective des investisseurs. Les résultats obtenus montrent une bonne capacité du modèle à **suivre la tendance générale du marché** et à **prédirer correctement la direction des mouvements de prix**, notamment pour des actions fortement influencées par l'activité sociale comme **GameStop**.

Un **tableau de bord interactif** développé avec **Streamlit** permet de visualiser les données boursières, le sentiment Reddit et les prédictions de manière claire et intuitive. Enfin, une analyse critique des résultats met en évidence les **limites du système** ainsi que plusieurs **perspectives d'amélioration**, telles que l'intégration de nouvelles sources de données et l'utilisation de modèles plus avancés.