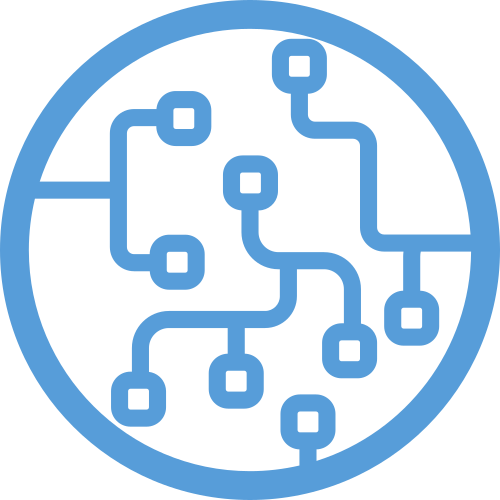
**Financial Sentiments**

****

***Étudiant : Capocasale Romain***

***Superviseurs : Ghorbel Hatem & Costa Pedro***

***Classe : INF3dlm-A***

***École : He-Arc, Suisse***

***Années : 2019-2020***

***Numéro de travail : 206***

***Abstract***

Dans le cadre du machine learning, le domaine de l'analyse de sentiments dans les textes a pris beaucoup d'intérêt ces dernières années grâce au succès et aux bonnes performances des algorithmes de classification de texte. L’objectif du projet est d’appliquer cette technologie d’analyse de sentiment au domaine de la finance dans le but de dire si un texte a plutôt un sentiment positif ou négatif.

Ce rapport retrace les différentes étapes de création de modèles de machine learning et de deep learning pour la prédiction de sentiment sur des phrases dans un contexte financier. Les modèles seront entraînés avec un vocabulaire propre à la financer. Un jeu de données initiales contenant plus de 1800 titres d’articles financiers m’a été fourni pour entraîner des modèles. Différentes étapes telles que le prétraitement des phrases, la représentation d’une phrase en nombre ont été effectués. Plusieurs types de modèles de machine learning et de deep learning ont été créés pour résoudre ce problème.

Des titres d’articles financiers étiquetés par des experts m’ont été fournis pour valider mes modèles.

Les résultats obtenus montrent qu’avec les techniques utilisées et le nombre de données à ma disposition, il n’est pas possible d’obtenir des résultats satisfaisants sur les données de validation. Les modèles développés n’ont pas de meilleurs résultats que des analyseurs de sentiments présents sur le web. Pour obtenir de meilleurs scores, il faudrait disposer d’un plus grand jeu de données.

Table des matières

[1 Introduction 4](#_Toc30192556)

[2 Cahier des charges 5](#_Toc30192557)

[3 Planning 6](#_Toc30192558)

[3.1 Planning initial 6](#_Toc30192559)

[3.2 Planning final 7](#_Toc30192560)

[3.3 Constatation sur le planning 8](#_Toc30192561)

[4 Analyse 8](#_Toc30192562)

[4.1 Etat de la technologie 8](#_Toc30192563)

[4.2 Schéma du pipeline des processus 8](#_Toc30192564)

[5 Métrique d’évaluation des modèles 9](#_Toc30192565)

[5.1 Métrique de régression 9](#_Toc30192566)

[5.1.1 RMSE 9](#_Toc30192567)

[5.1.2 MAE 9](#_Toc30192568)

[5.2 Métrique de classification 9](#_Toc30192569)

[5.2.1 La précision 10](#_Toc30192570)

[5.2.2 Le rappel 10](#_Toc30192571)

[5.2.3 L’exactitude 10](#_Toc30192572)

[6 Jeu de données 10](#_Toc30192573)

[6.1 Titres d’articles financier 10](#_Toc30192574)

[6.1.1 Analyse des données 11](#_Toc30192575)

[6.2 Tweets financiers 11](#_Toc30192576)

[6.2.1 Analyse des données 12](#_Toc30192577)

[6.3 Tweets quelconques 13](#_Toc30192578)

[6.4 Titres d’articles financiers annotés par des experts 13](#_Toc30192579)

[7 Analyseur de sentiment présent sur le web 14](#_Toc30192580)

[8 Solutions développées 15](#_Toc30192581)

[8.1 Prétraitement des données 15](#_Toc30192582)

[8.1.1 Traitement du texte 15](#_Toc30192583)

[8.2 Représentation d’une phrase 16](#_Toc30192584)

[8.2.1 TF-Idf 16](#_Toc30192585)

[8.2.2 Word2vec 17](#_Toc30192586)

[8.2.3 Padding 21](#_Toc30192587)

[8.3 Modèle de machine learning 21](#_Toc30192588)

[8.3.1 Modèle de régression 21](#_Toc30192589)

[8.3.2 Modèle de classfication 21](#_Toc30192590)

[8.4 Modèle de deep learning 22](#_Toc30192591)

[8.5 Évaluation des modèles de machines learning 22](#_Toc30192592)

[8.5.1 Titres d’articles financiers 22](#_Toc30192593)

[8.5.2 Tweets financiers 26](#_Toc30192594)

[8.5.3 Tweets quelconques 29](#_Toc30192595)

[8.6 Évaluation des modèles de deep learning 29](#_Toc30192596)

[8.6.1 LSTM avec padding 29](#_Toc30192597)

[8.6.2 GRU avec padding 30](#_Toc30192598)

[8.6.3 CNN avec Word2Vec de Glove 31](#_Toc30192599)

[8.6.4 RNN avec Word2vec financier 32](#_Toc30192600)

[8.7 Conclusion des solutions développés 33](#_Toc30192601)

[9 Approfondissement possible 34](#_Toc30192602)

[9.1 Utilisation de Bert 34](#_Toc30192603)

[9.2 Agrandissement du jeu de données 34](#_Toc30192604)

[10 Conclusion 35](#_Toc30192605)

[11 Bibliographie 35](#_Toc30192606)

[11.1 Internet 35](#_Toc30192607)

[11.2 Livres 36](#_Toc30192608)

[12 Annexes 36](#_Toc30192609)

[12.1 Journal de bord 36](#_Toc30192610)

# Introduction

Dans le cadre du 5ème semestre de Bachelor dans la filière développement logiciel et multimédia, nous avons pour objectif de réaliser un projet individuel qui permet de nous préparer le mieux possible au travail de Bachelor. J’ai alors sélectionné un projet dans le domaine du machine learning.

Dans le cadre du machine learning, le domaine de l'analyse de sentiments dans les textes a pris beaucoup d'intérêt ces dernières années grâce au succès et aux bonnes performances des algorithmes de classification de texte. Dans un premier temps cette technologie a été utilisée pour analyser les sentiments dans des domaines tels que les commentaires de films, les hôtels, les restaurants, etc.

Le but du projet est d’appliquer cette technologie d’analyse de sentiment au domaine du marché financier dans le but de dire si un texte à plutôt un sentiment positif ou négatif. Ce sentiment pourrait alors être utilisé dans le domaine du FOREX (Foreign Exchange Market). Une information perçue comme positive pourrait indiquer une montée d’une devise par rapport à une autre tandis qu’une information négative pourrait en indiquer la chute.

L’objectif du projet est de créer un analyseur de sentiment pour le domaine de la finance. Le modèle sera entraîné avec des textes et vocabulaires propres au thème de la finance. Dans un premier temps, les solutions d’analyse de sentiment présente sur le web seront analysées pour résoudre ce problème. Dans un second temps, le but est de créer différents modèles de machine learning dans le but de prédire un sentiment par rapport à un texte financier.

# Cahier des charges

Le but du projet est de créer un modèle d'analyse de sentiment pour le domaine de la finance. Un analyseur de sentiment est un programme informatique qui en entrée demande une phrase ou un texte et en sortie indique un sentiment pour cette phrase sur une échelle de -1 à 1. Une valeur de -1 pour un sentiment négatif et 1 pour un sentiment positif. L’intervalle peut être discret ou continue.

Étant donné que le domaine de la finance dispose de son propre vocabulaire, il est intéressant de développer un modèle spécialisé dans ce secteur dans le but d'obtenir de meilleur résultat qu'un analyseur de sentiment global qui aurait été entraîné à faire des prédictions sur tout le vocabulaire de la langue anglaise.

Pour l'entraînement et le test du modèle, plusieurs sources de donnée différentes ont été fournies notamment, un ensemble de titres d'article de finance avec la valeur du sentiment correspondant. Les modèles seront réalisés en Python.

Pour ce projet, il faudra dans un premier temps, comparer le score des analyseurs de sentiment existant sur le web avec les jeux de donnée fourni. Par la suite, le but est de développer un modèle spécialisé dans l'analyse de sentiment pour la finance. La première étape consistera à prétraiter la phrase pour enlever la ponctuation, les déterminants, etc. Il faudra par la suite, transformer la phrase en vecteur dans le but d'être utilisé dans un modèle de machine learning. Différents modèles de machine learning devront être implémentés et testés à l'aide des jeux de donnée fournie. Par la suite, un modèle plus complexe utilisant des techniques de deep learning pourra être implémenté. Il faudra également comparer ces différents modèles entre eux pour sélectionner celui effectuant le moins d'erreurs de prédiction.

Le but final (hors projet), serait d'utiliser ce modèle d'analyse de sentiment dans le but de suivre la tendance haussière et baissière dans les pairs de devises (par exemple USD/CHF et EUR/USD). Des sources textuelles seraient récoltées depuis le site de Reuters et d'autres canaux financiers spécialisés dans le but d'en prédire un sentiment (entre -1 et 1) d'une période de trading (une heure, une demi-journée, journée, semaine ...) et relatif à un sujet (par exemple Brexit, échanges commerciaux avec la Chine ....).

Ce classificateur avec d'autres en cours de développement, sera intégré dans un outil de prédiction du taux change.

# Planning

## Planning initial

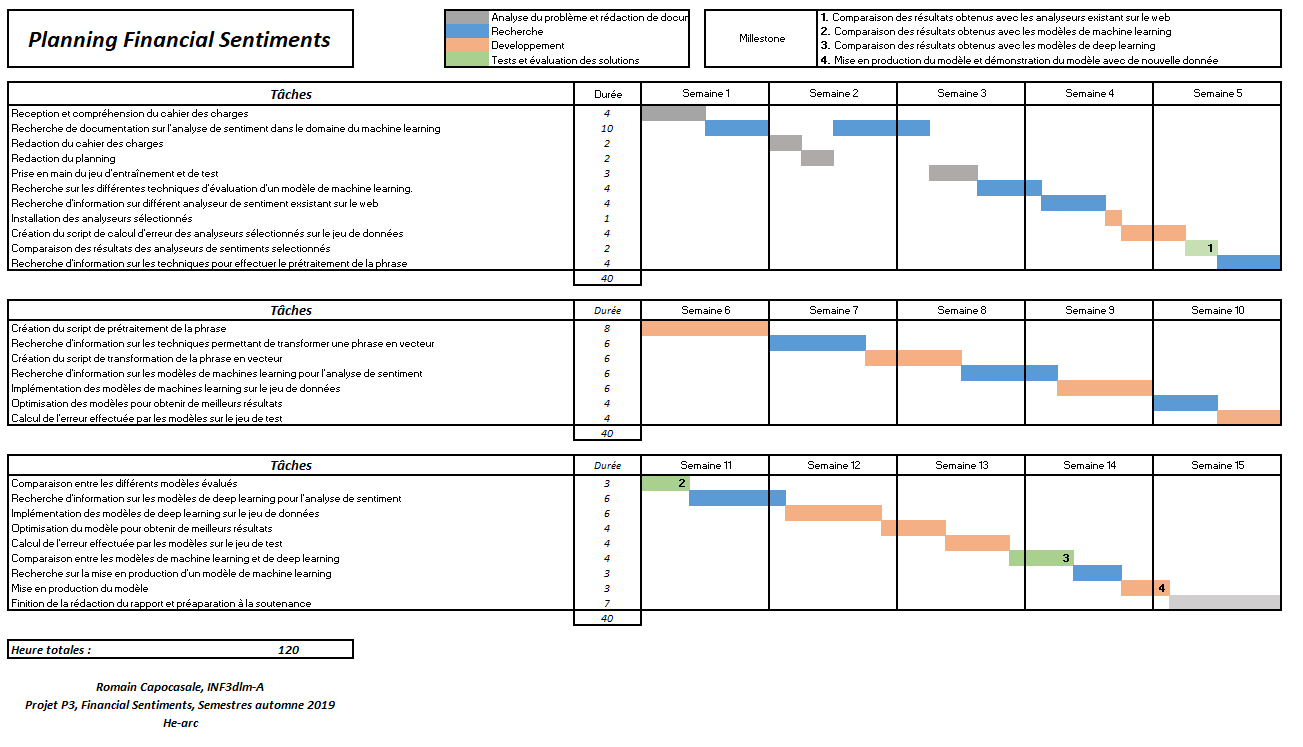


Figure - Planning initial

## Planning final

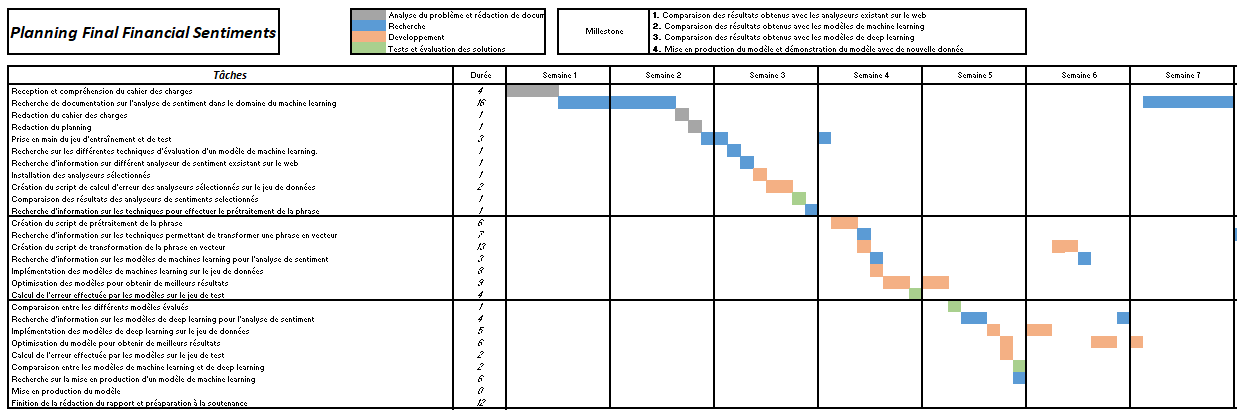


Figure - Planning final (1)

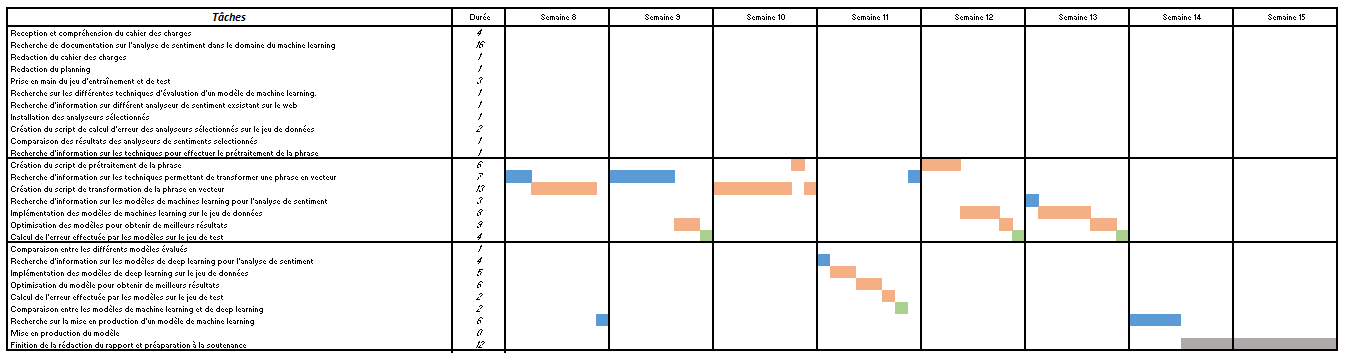


Figure - Planning final (2)

## Constatation sur le planning

On peut remarquer que les étapes du début de projet ont été réalisées plus rapidement que prévu. Comme j’ai reçu des jeux de données à différents moments du projet, on constate que j’ai recommencé plusieurs fois les étapes d’analyse des données, de prétraitement, de représentation de la phrase et de création des modèles.

# Analyse

## État de la technologie

Il existe de nombreux analyseurs de sentiment sur le web comme par exemple : Textblob, VADER (nltk), etc. Ces analyseurs sont très simples à utiliser, il suffit d’indiquer la phrase et la librairie va indiquer la polarité de la phrase entre -1 et 1. Ces analyseurs ont cependant été entraînés sur tout le vocabulaire de langue anglaise et non sur le vocabulaire de la finance spécifiquement. Ils auront alors tendance à commettre plus d’erreurs qu’un classificateur entraîné sur un vocabulaire spécifique. Un modèle entraîné avec un vocabulaire spécifique donnera une pondération plus importante aux mots qu’il considère plus importants dans le contexte donné.

Si on regarde dans le domaine de la finance spécifiquement, d’après mes recherches ils n’existent pas de libraire ou d’analyseur de sentiment spécifique à la finance. On ne trouve sinon pas mal d’articles qui expliquent l’utilité du NLP (Natural Language Processing) appliqué au domaine de la finance.

Avec ces différentes recherches, on constate donc qu’il serait intéressant de créer un modèle de Machine Learning entraîné avec un vocabulaire spécifique à la finance.

## Schéma du pipeline des processus

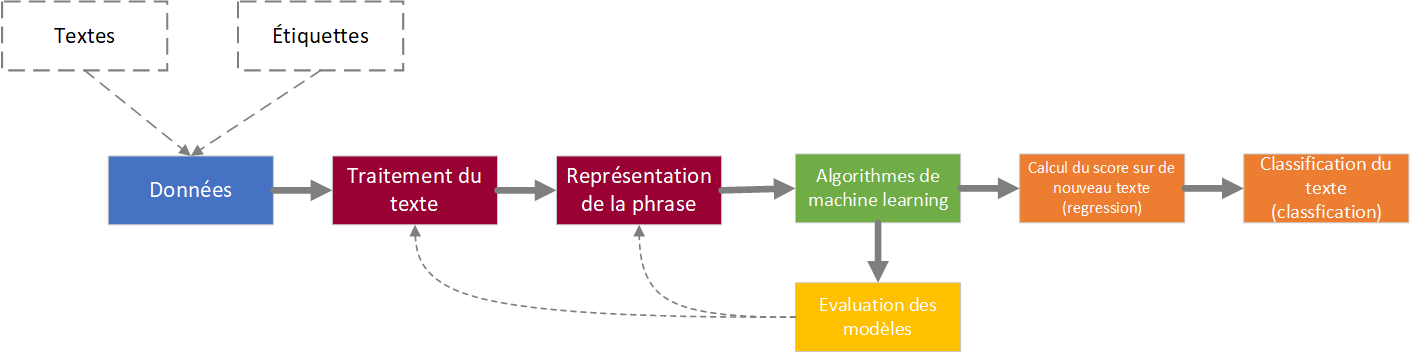


Figure - Machine Learning pipeline

Ce schéma montre les différentes étapes qui seront effectuées dans ce projet dans le but d’obtenir un modèle de machine learning.

Dans un premier temps, les données provenant de différentes sources devront être recueillies puis séparées entre les textes (phrase à analyser) et les étiquettes (sentiment pour une phrase). C’est également dans cette étape que les données devront être analysé dans le but d’en ressortir les tendances, subtilité, etc.

Dans un second temps, les données doivent être prétraitées, dans le but de pouvoir être introduites dans un algorithme de machine learning. Le texte doit d’abord être traité dans le but de retirer le bruit. C’est-à-dire enlever la ponctuation, enlever certains mots, etc.

Après avoir traité le texte, il faut maintenant trouver une manière de représenter les phrases sous forme de nombre. En effet, les algorithmes de machine learning ne peuvent travailler qu’avec des nombres et non du texte. Différentes techniques comme le word2vec et le tfidf pourront être utilisés pour cette étape.

À présent, différents algorithmes de machines learning doivent être étudiés et implémentés pour résoudre le problème. Les données seront alors séparées en données d’entraînement et de test. Le modèle sera alors conçu à partir du jeu d’entraînement.

Une fois le modèle conçu, il faut l’évaluer avec les données de test et calculer différents scores qui permettront de se faire une idée de la fiabilité ou non du modèle. Selon les scores obtenus, il peut être intéressant de modifier certains paramètres des étapes vues précédemment dans le but d’améliorer les modèles. Certains paramètres du modèle (hyperparamètres) peuvent aussi être adaptés et modifiés.

Une fois qu’un modèle satisfaisant a été obtenu, il est maintenant possible d’effectuer de nouvelle prédiction sur de nouveaux textes. Le modèle prédira un score entre -1 et 1 sur un intervalle continu pour ce texte (régression) ou sur un intervalle discret (classification).

# Métrique d’évaluation des modèles

Avant de commencer à parler des solutions développées, il est important de définir les métriques qui nous permettront d’évaluer notre modèle. C’est grâce à cela, qu’on pourra dire si un modèle est viable ou non.

## Métrique de régression

### RMSE

* La RMSE (Root Mean Square Error) est la racine carrée de l’erreur quadratique moyenne, elle permet de donner une idée de l’importance des erreurs de prédictions commises par le système.

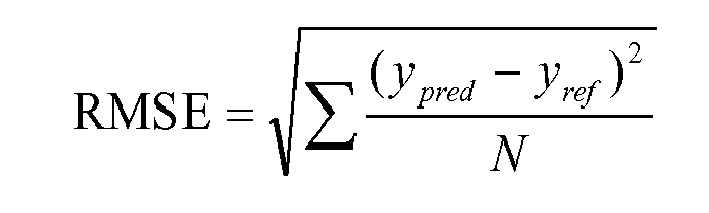


Figure - RMSE formule

### MAE

* La MAE (Mean Absolute Error) est l’erreur absolue moyenne, elle permet de donner moins d’importance aux valeurs extrêmes.

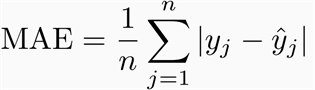


Figure - MAE formule

## Métrique de classification

Pour les métriques de classification il faut d’abord définir ce que sont :

* Les vrais négatifs (TN) : C’est le nombre de phrases négatives qui sont classées correctement dans la classe négative.
* Les faux positifs (FP) : C’est le nombre de phrases négatives qui sont classées incorrectement c’est-à-dire dans la classe positive.
* Les faux négatifs (FN) : C’est le nombre de phrases positives qui sont classées incorrectement c’est-à-dire dans la classe négative.
* Les vrais positifs (TP) : C’est le nombre de phrases positives qui sont classées correctement dans la classe positive.

### La précision

La formule de la précision est :

Figure - Précision formule

### Le rappel

La formule du rappel est :

Figure - Rappel formule

Pour ce projet il est important d’avoir un bon compromis entre la précision et le rappel. C’est-à-dire que l’on souhaite que la précision et le rappel ai un peu près la même valeur. Il faut un bon compromis, car il est autant important de repérer les phrases négatives que positif.

### L’exactitude

L’exactitude en anglais « accuracy » correspond au nombre de prédictions correctes divisé par le nombre total d’observations.

# Jeu de données

## Titres d’articles financiers

Deux sources différentes de données m’ont été fournies[[1]](#footnote-1) [[2]](#footnote-2). C’est le premier jeu de données que j’ai eu ma disposition pour ce projet. Les fichiers étaient au format JSON. Les 2 jeux de données contenaient chacun une colonne avec la phrase et le sentiment correspondant. Cependant, chaque jeu de données avait des colonnes supplémentaires que l’autre jeu de donnée n’avait pas avec d’autres informations. Un contenait par exemple l’entreprise qui était le sujet du texte et dans l’autre jeu de données une colonne contenait la provenance des textes. Même si ces informations auraient pu être intéressantes à inclure dans le modèle, j’ai préféré ne garder que les colonnes avec la phrase et le sentiment correspondant dans le but d’avoir exactement les mêmes colonnes pour les deux jeux de données. Cela me permet d’accroître la taille du jeu de données final. De plus, si d’autres variables avaient été incluses dans le modèle, il aurait fallu les renseigner lors de future prédiction ce qui peut être contraignant.

J’ai alors créé un script python (présent sous dataset/headlines\_dataset/rebuild\_data.py) permettant de récupérer les deux jeux de donnés et d’en ressortir un unique avec seulement les informations sélectionnées précédemment. Les données retournées sont sous forme d’un fichier au format JSON.

### Analyse des données

Une fois les données récupérées, il est important d’effectuer une analyse des données dans le but de voir différentes statistiques sur les données, des tendances ou des subtilités. Ceci dans le but de bien comprend comment elles sont structurées, afin de construire un modèle efficace. Il est important de noter que les statistiques ci-dessous concernent les données non prétraitées.

On remarque alors différentes statistiques sur les données, voici les plus importantes (pour plus de statistiques, consulter le notebook) :

* Il y a en tout 2866 lignes (2866 phrases et 2866 sentiments correspondants). On remarque qu’il y a assez peu de données.
* Il y a 1756 phrases positives, 1045 phrases négatives et 65 phrases neutres. On remarque alors qu’il y a plus de phrases positives que négatives.
* Il y a environ 8 mots par phrase. Ce qui est une longueur courte pour une phrase.

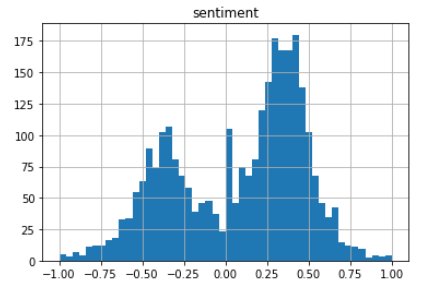
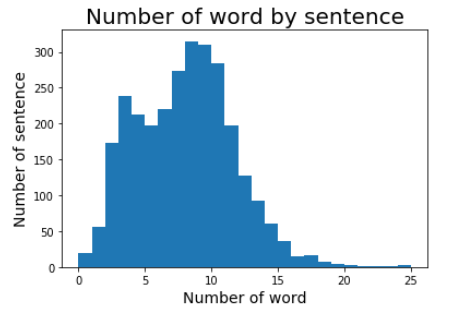


Figure – Nombre de phrases par polarité (titres articles financiers)

Figure - Nombre de mots par phrase (titres articles financiers)

Sur la figure 9 un histogramme avec sur l’axe X, la polarité d’une phrase et sur l’axe Y la quantité de phrases. Sur la figure 10 un graphe du nombre de mots par phrase.

## Tweets financiers

Mon second jeu de données provient du dépôt GitHub suivant : <https://github.com/AndMu/Market-Wisdom>. J’ai reçu ce jeu de données à la 12ème semaine du projet. Ces données ont été récoltées par un robot sur Twitter. Le robot avait pour but d’analyser les tweets financiers et de compter le nombre d’adjectifs dans les catégories suivantes :

* Anger -> négatif
* Anticipation -> neutre
* Disgust -> négatif
* Fear -> négatif
* Joy -> positif
* Sadness -> négatif
* Surprise -> neutre
* Trust -> positif

Ces différentes données proviennent de plusieurs fichiers csv différents. Il a d’abord fallu remettre toutes ces différentes données ensemble puis faire une moyenne pondérée entre avec les différents adjectifs. Les adjectifs positifs avec une pondération de 1 et -1 pour les négatifs. Par exemple, si une phrase contient 2 adjectifs de type Trust, 1 de type Anger et 3 de type Sadness, la polarité du tweet correspondant sera de (2 + (-4)) / 6 = -0.33.

Ces différentes opérations se trouvent dans le notebook : « *MarketDataPreProcess.ipynb »* dans le dossier *« dataset/market\_dataset/ ».*

### Analyse des données

Ce jeu de données est de taille importante. Il contient plus de 1,2 million de tweets. Voici différentes statistiques sur ces données :

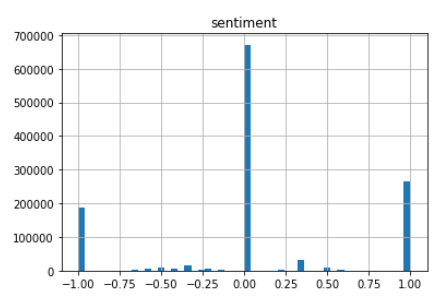
* Il y a 315'852 phrases positives, 238'591 phrases positives et 671'275 phrases neutres.
  + On remarque qu’il y a énormément de phrases neutres. Les phrases neutres n’ont pas beaucoup d’utilité pour la construction d’un modèle. La moitié des phrases du jeu de données n’ont donc pas beaucoup d’utilité. De plus, il a de grandes chances pour que quand la phrase à une polarité de 0, il s’agit juste du robot qui a scanné les tweets qui n’as pas réussi à déterminer les sentiments du tweet.
  + On constate de nouveau qu’il y a plus de phrases négatives que positives
* Il y a moyenne 13 mots par phrases (plus du double par rapport au premier jeu de données).

Figure - Nombre de phrases par polarité (Tweets financiers)

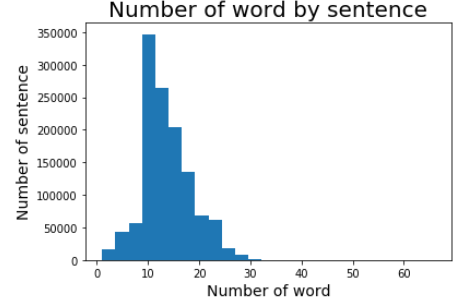


Figure - Nombre de mots par phrase (Tweets financiers)

Sur la figure 11 un histogramme avec sur l’axe X, la polarité d’une phrase et sur l’axe Y la quantité de phrases. Sur la figure 12 un graphe du nombre de mots par phrase. On constate qu’il y a comme indiqué précédemment beaucoup de phrases avec une polarité de 0 dans le jeu de données. On remarque sinon que le sentiment est sois -1 sois 1. Il y a très peu de valeur intermédiaire.

Pour pouvoir utiliser ce jeu de données de manière efficace, il faudra donc supprimer une grande quantité de phrases qui ont une polarité de 0. J’ai décidé d’en supprimer les 3/4. Comme il y a énormément de données, je n’en ai utilisé qu’une petite partie sélectionnée aléatoirement (30'000) pour que le temps d’entraînement des modèles ne soit pas trop long.

## Tweets quelconques

Ce jeu de donnée provient de <https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/> . Il m’a été fourni à la semaine 7 du projet. Il contenait 160'000 tweets annotés dans n’importe quelle catégorie. C’est-à-dire pas forcément financiers. De plus, ces tweets étaient annotés à partir des smileys présents dans le texte. Ce qui ne permet pas de définir la polarité de manière précise et sûre.

Une liste de différents topiques financiers m’a été fournie dans le but d’extraire les tweets contenant au moins un de ces mots. Après extraction, il ne restait que 9'000 tweets. Après analyse des tweets restants, j’ai pu constater que les qualités de ce ses phrases étaient médiocre. En effet, elle contenait certes des mots financiers, mais le contexte n’avait dans la grande majorité rien à voir avec la finance.

Voici certains tweets obtenus après extraction :

* Oracle buying Sun for $7.4 billion..not a happy day
* I missed out on the $3 off Huggies coupons!
* Another day another dollar... i don't want to work.
* I want a subscription to New Scientist magazine. Too bad it costs $72 a year.
* I have headache =$ I still have to go to school tough.

On constate donc que la plupart des phrases n’ont aucun rapport avec la finance sauf quelques exceptions comme la première phrase (il m’a quand même fallu pas mal de temps pour la trouver dans l’ensemble des tweets). On voit que les phrases trouvées sont loin de titres d’articles financiers. C’est pour cette raison que cette solution a été abandonnée.

Le script de prétraitement des tweets peut être trouvé dans le dossier : *dataset/tweets\_datatset/rebuild\_data.ipynb.* Un notebook contenant différentes étapes pour la création de modèles à partir de ce jeu de données se nomme *TweetsMachineLearning.ipynb*.

Ces modèles donnant de mauvais résultats en grande partie du a la qualité des données en entrée, ces modèles ont été abandonnés et ne seront pas détaillés dans ce rapport. Cependant la marche à suivre peut se trouver dans le notebook cité ci-dessus.

## Titres d’articles financiers annotés par des experts

Ces données proviennent de l’entreprise qui a mandaté le projet. Elles ont été annotées à la main par des experts financiers. Ce jeu de données contient 200 phrases. Ce sont donc ces données qui devront être utilisées comme tests pour valider les différents modèles. Ces données m’ont été fournies à la semaine 14 du projet. Le problème avec ces données est qu’elles n’ont pas été annotées à partir des mots présents dans la phrase, mais à partir de la montée du dollar par rapport à l’euro et inversement. C’est-à-dire qu’une phrase peut être remplie de mot négatif, mais qu’elle sera annotée positivement parce que dans le contexte de la phrase l’euro augmente.

Voici par exemple 4 phrases du jeu de données de test

* « *berlin - uk foreign secretary hunt says we all want to work with urgency to avoid brexit going wrong* » annoté positivement
* « *snb's zurbruegg says situation on forex market remains fragile* » annoté positivement
* « *offshore chinese yuan vs dollar rises above 6.8 for the first time since july 2017 <cnh=ebs>*» annoté négativement
* « *draghi says rate guidance very effective as reflected by surveys and markets* » annoté avec une polarité neutre

On remarque par exemple dans la première phrase que les mots « urgency », « avoid » et « wrong » ainsi que la phrase ont un contexte négatif pourtant la phrase est annotée positivement.

Pour la seconde phrase le contexte de la phrase et le mot « fragile » pourrait présager que la phrase est négative pourtant elle est annotée positivement.

Le contexte de la troisième phrase ainsi que le mot « rises » ont un contexte positif, mais la phrase est annotée de manière négative.

Pour la dernière phrase, on constate que le contexte de la phrase ainsi que le mot « very » et « effective » ont un contexte positif, mais la phrase est annotée avec une polarité neutre.

On remarque donc que ce jeu de données est annoté de manière différente du premier jeu de données qui m’a été fourni.

# Analyseur de sentiment présent sur le web

Pour le projet, il m’a été demandé de comparer les scores d’analyseur de sentiment présent sur le web. Les scores obtenus ici seront utilisés comme score de référence pour l'évaluation des futurs modèles. Le but des futurs modèles est de battre ces scores. Sur la base des recherches effectuées, deux modèles d’analyse de sentiment ont été retenus : la librairie Textblob et VADER un sous package de la libraire NLTK. J’ai calculé ces scores avec le jeu de données des titres d’articles financiers.

Voici les scores qu’obtiennent ces deux analyseurs avec les titres d’articles financiers:

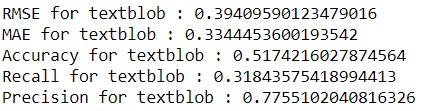
* Pour Textblob :

Figure - Score Textblob headlines

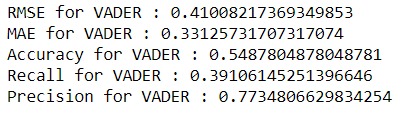
* Pour VADER :

Figure - Score Vader headlines

On constate que les résultats obtenus sont plutôt mauvais. Les 2 modèles obtiennent une précision d’à peine plus de 50%, ce qui correspond à un choix presque aléatoire de la classe. J’ai également constaté que ces 2 analyseurs ont souvent tendance à prédire un sentiment de 0 pour les phrases.

Voici les scores que j’obtiens pour le jeu de données des titres d’articles financiers annotés par les experts :

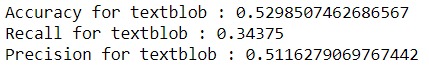
* Pour Textblob :

Figure - Score Textblob headlines expert

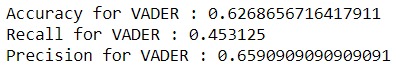
* Pour Vader :

Figure - Score Vader headlines expert

On remarque encore une fois que les scores obtenus sont à peine au-dessus des 50% d’exactitude.

Cependant comme à l’étape précédent, les deux analyseurs ont tendance à prédire beaucoup de 0 (environ la moitié des prédictions sont des 0). Cela pose un problème, car la polarité de 0 ne peut pas être considérée comme positive ni négative. Il faudra alors sois supprimer toutes les prédictions qui ont une probabilité de 0, sois classes les phrases avec une polarité de 0 dans la classe positive ou négative. Pour les scores présentés précédemment j’ai choisi de considérer le 0 comme négatif ce qui va donc biaiser les résultats.

Au vu des différents scores, on constate qu’il est très intéressant de construire un modèle de machine learning avec pour but de battre les scores ci-dessus.

# Solutions développées

Voici les différentes solutions développées ainsi que l’argumentation sur le choix de ces solutions sur la base des recherches effectuées en amont sur le domaine.

## Prétraitement des données

Ici, le but est de prétraiter les données dans le but de pouvoir les introduire dans un algorithme de machine Learning ou de deep learning.

### Traitement du texte

#### Spacy

Voici les différents processus qui ont été effectués pour prétraiter la phrase :

* **Tokenisation** : Cette étape consiste à découper la phrase en mot.
* **Suppression des stop word** : Cette étape consiste à supprimer des mots qui ont peu d’importance dans la phrase comme « the », « are », etc.
* **Tagging et suppression de certains mots tagués** : Le tagging consiste à détecter à quelles classes appartient un mot comme par exemple nom, adjectif, verbe, etc. Par la suite, il faut supprimer le type de mot que l'on ne veut pas garder dans la phrase. Ceci peut être utile pour enlever par exemple les déterminants.
* **Suppression des mots en dessous d’une certaine taille** : Ce processus consiste à ne garder que les mots d'une certaine taille dans la phrase.
* **Suppression de la ponctuation** : Cette étape consiste à supprimer les caractères comme « ! », « ? », « , » etc.
* **Lemmatisation** : Ce processus consiste à ne garder que la racine des mots. Par exemple, les mots comme « runs », « running » seront transformés en « run ». Ceci permet de réduire le nombre de mots dans le dictionnaire final.

Il est à noter que les différents processus ci-dessus ne sont pas tous effectués systématiquement lors du traitement de la phrase. Le but est de regarder quels processus font améliorer ou non le score du modèle.

Le traitement du texte est effectué avec la librairie Spacy[[3]](#footnote-3). Pour effectuer ceci, j’ai créé une classe nommée « TextProcessor » héritant de « BaseEstimator » et de « TransformerMixin ». Le fait d'hériter de ces 2 classes permettra d'inclure cette classe dans le pipeline sklearn. Les différents processus vus ci-dessus peuvent être ajoutés ou non au traitement via le constructeur de la classe. Le principal avantage de cette méthode est qu'elle se combine avec l'utilisation de la classe GridSearchCV() ou RandomizedSearchCV() qui permet de chercher les meilleurs hyperparamètres pour un modèle. Cette classe se trouve dans le dossier *tools/TextProcessor.py*.

#### Gensim

La bibliothèque python Gensim possède une fonction permettant d’effectuer le prétraitement d’une phrase[[4]](#footnote-4). Cette fonction met les mots en minuscule, les tokenize, enlève les accents et par défaut enlève les mots plus petits que 2 caractères et plus grands que 15 caractères. Cette méthode est moins complète que celle-ci dessus, mais à l’avantage d’être extrêmement rapide.

## Représentation d’une phrase

### TF-Idf

Comme un modèle de machine Learning fonctionne avec des nombres et non du texte, il faut trouver un moyen de représenter les phrases sous forme de nombres. Le TF-idf permet d’effectuer cette opération en représentant un mot sous forme de nombre.

Le TF-idf se compose de 2 parties, le TF (Term Frequency) qui est le nombre de fois qu’apparaît le mot dans la phrase. Cette fréquence peut être normalisée à l’aide d’un logarithme.

Où ft,d représente la fréquence du mot dans la pharse.

L’idf (Inverse Document Frequency) est le nombre total de documents dans le corpus divisé par le nombre de documents où le terme apparaît.

Figure - Tf formule

Figure - Idf formule

Où D est le nombre total de documents dans le corpus et dj le nombre de documents où le terme apparaît.

Pour finir le score, TF et idf sont multipliés. Ce score est calculé pour chaque phrase et représenté dans une matrice sparse[[5]](#footnote-5) de scipy. C’est cette matrice qui est passée au différent modèle.

Cette technique a été utilisée dans mon projet pour les modèles de machines learning.

### Word2vec

Trois modèles de Word2vec ont été utilisés pour représenter une phrase dans l’espace. En premier le word2vec de la librairie Gensim[[6]](#footnote-6). En deuxième, un word2vec financier qui m’a été fourni par un assistant de l’équipe de recherche. En troisième, le word2vec de Glove[[7]](#footnote-7).

#### Gensim

La méthode utilisée via Gensim pour représenter la phrase dans l’espace n’est pas un word2vec, mais un doc2vec[[8]](#footnote-8). Un doc2vec permet de représenter une phrase entière dans l’espace contrairement au word2vec qui permet de représenter seulement un mot. La classe doc2vec utilisée est un wrapper sklearn pour Gensim. Ce qui signifie que le doc2vec peut donc être directement utilisé avec sklearn et par exemple être placé dans un pipeline.

Le doc2vec représente une phrase avec 100 composantes. Pour représenter les phrases dans l’espace il suffit d’utiliser la méthode fit\_transform() de l’objet doc2vec. Les phrases peuvent être après envoyées directement au modèle.

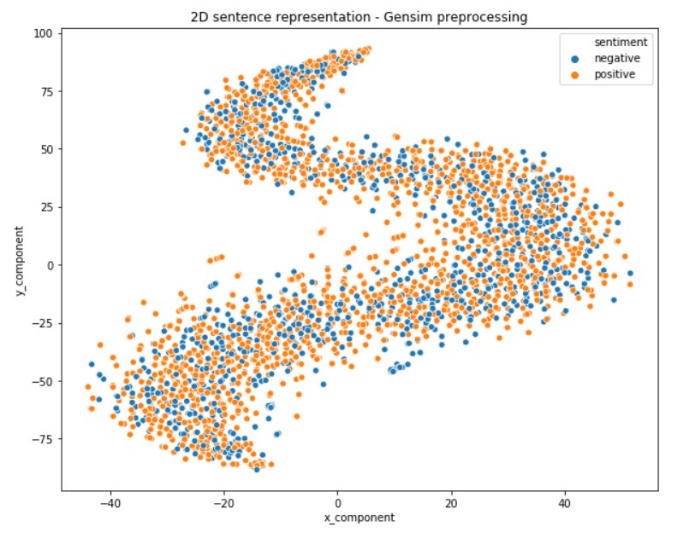
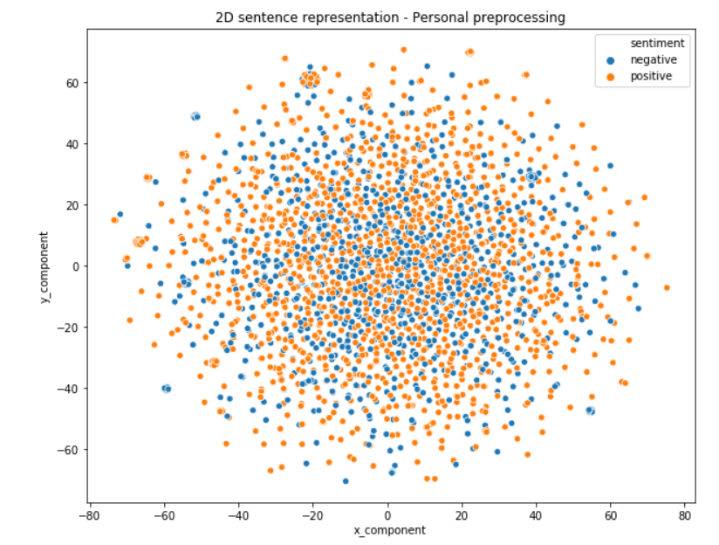
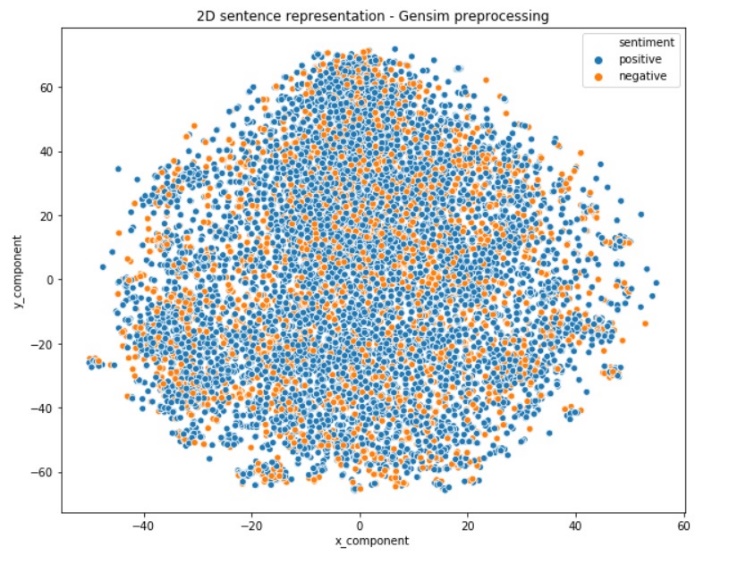
Voici une représentation T-sne[[9]](#footnote-9) du jeu des données des HeadLines financière. Une représentation T-sne permet de réduire le nombre de dimensions d’un vecteur donné. Il est utile pour visualiser les données en 2 dimensions. Comme les vecteurs ont ici deux dimensions, il peut être utile de les représenter en 2 dimensions pour par exemple voir si les phrases sont bien réparties dans l’espace, si les phrases positives sont proches des positives et inversement, etc. …

Figure - T-sne gensim doc2vec

On peut voir le jeu de données des headlines financières représenté dans l’espace. À gauche, avec le prétraitement effectué avec Gensim et à droite avec le prétraitement effectué par Spacy. Avec Gensim, on peut voir que les phrases sont moins éparpillées dans l’espace qu’avec l’autre méthode, même si on peut voir un amas de phrases dans le coin en bas à gauche de la figure gauche. On peut constater dans les 2 cas que les phrases positives ne sont pas proches des phrases positives et inversement pour les négatives. Le modèle de machine learning aura donc beaucoup de mal à séparer les phrases positives et négatives.

Figure - T-sne spacy doc2vec



On peut voir ici une dernière représentation T-sne mais cette fois si avec le modèle des tweets financiers. Seules 10'000 phrases ont été représentées sur 1'000'000 (la fonction T-sne demande beaucoup de temps de calcul). On constate également qu’il n’a pas de séparation claire entre les phrases positives et négatives. Le modèle aura donc du mal à généraliser les résultats.

Cette technique de représentation des phrases avec Gensim est utilisée dans mon projet pour des modèles de machines learning.

#### Word2vec financiers

Le word2vec financier fourni par un assistant de l’équipe de recherche a été entraîné sur un million de mots financiers. Un mot est représenté avec 100 composantes. Comme le word2vec fournit la représentation d’un mot et non d’une phrase dans l’espace, il faut utiliser un système de padding (expliqué plus bas) avec la phrase la plus longue du jeu de données. En effet, il faudra donner des vecteurs de même taille au modèle sklearn. De plus, les mots non présents dans le jeu de donnée seront représentés par un vecteur contenant des 0.

Différentes méthodes de représentation dans la phrase ont été développées :

* Concaténation des vecteurs : Les vecteurs représentant les mots sont concaténés les uns à la suite des autres. Si la phrase la plus longue du jeu de données à une taille de 25, et que la phrase à représenter à une taille de 14, il faudra concaténer 11 vecteurs remplis de 0 à la fin du vecteur (c’est le principe du padding). Si les vecteurs représentant un mot ont 100 dimensions, la taille finale d’un vecteur pour une phrase sera de 25'000 composantes.
* Concaténation des vecteurs multipliés par le score TF-Idf du mot : C’est le même principe que la méthode précédente sauf qu’avant de concaténer le mot, on va multiplier chaque composante du vecteur par son score TF-Idf. Ceci permet également de prendre en compte la fréquence du mot dans le corpus et dans un document.
* Somme des vecteurs : Les composantes des vecteurs sont sommées avec le vecteur précédent. Peu importe la taille de la phrase, le vecteur aura toujours 100 composantes.
* Somme des vecteurs multiplié par le score Tf-Idf : C’est le même principe que l’étape précédente sauf qu’avant de sommer les composantes, on multiplie par le score TF-Idf du mot.
* Moyenne des vecteurs : Les composantes des vecteurs sont sommes et divisé par la longueur de la phrase.
* Moyenne des vecteurs multipliés par le score TF-Idf : C’est le même principe que l’étape précédente sauf qu’avant de sommer les composantes, on multiplie par le score TF-Idf du mot. Les composantes sont également divisées par la taille de la phrase.

Une fois que tous les vecteurs représentant une phrase font la même taille, ils peuvent être envoyés au modèle de machine learning.

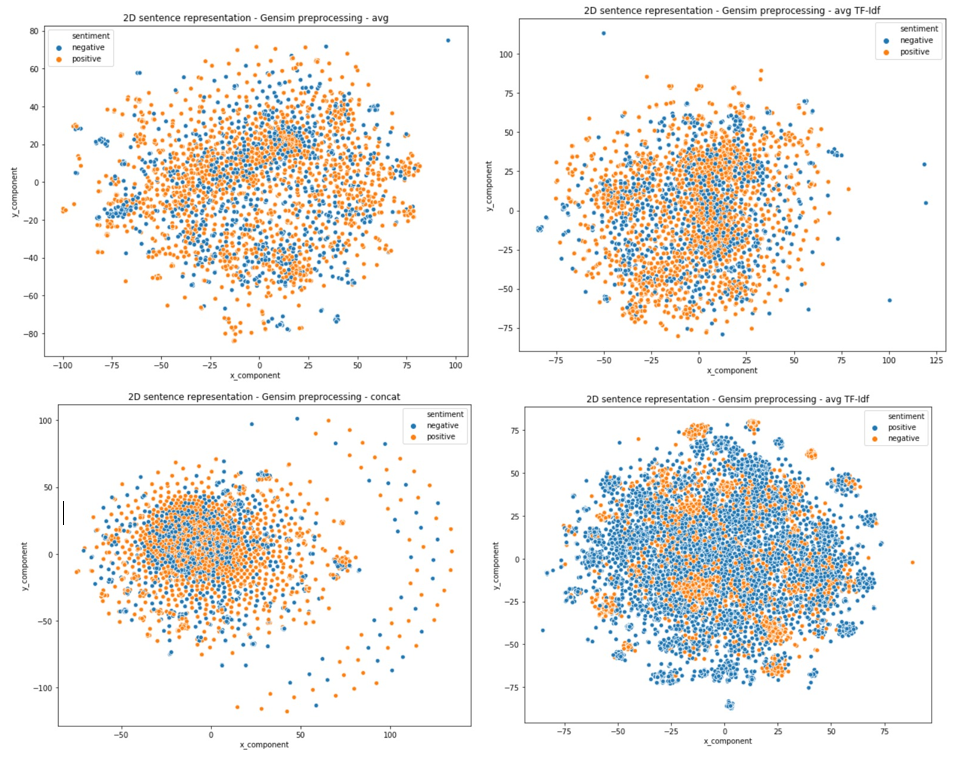


Figure - T-Sne word2vec financier

On retrouve ci-dessus 4 représentations T-sne. Les 3 graphes avec les moins de points sont les vecteurs appartenant au jeu de données des Headlines financières. Le graphe en bas à droite appartient aux tweets financiers. Différentes techniques ont été utilisées ici notamment :

* La moyenne : On constate que la moyenne est la représentation qui disperse le mieux les données. On constate que les phrases positives ont à peine plus tendance à se retrouver proches des phrases positives et inversement. La démarcation entre positif et négatif n’est cependant pas flagrante.
* La moyenne multipliée par le TF-Idf : Là encore une fois les phrases positives ont tendance à être positionnées vers les positives et inversement. Mais la démarcation n’est pas flagrante.
* La concaténation : On constate que cette méthode regroupe beaucoup les données et forme des pattern bizarres dans le bord droit. On constate encore une fois qu’il n’y a pas de séparation entre les phrases positives et négatives.

Il est bien sûr possible de créer plein d’autres exemples de représentation T-sne. Notamment, en modifiant le jeu de données utilisé, la méthode de prétraitement du texte et la représentation en vecteurs. Ci-dessus, ne sont affichés que 4 exemples. Cependant avec tous les différents exemples que j’ai testés aucun ne séparait clairement les phrases positives des phrases négatives.

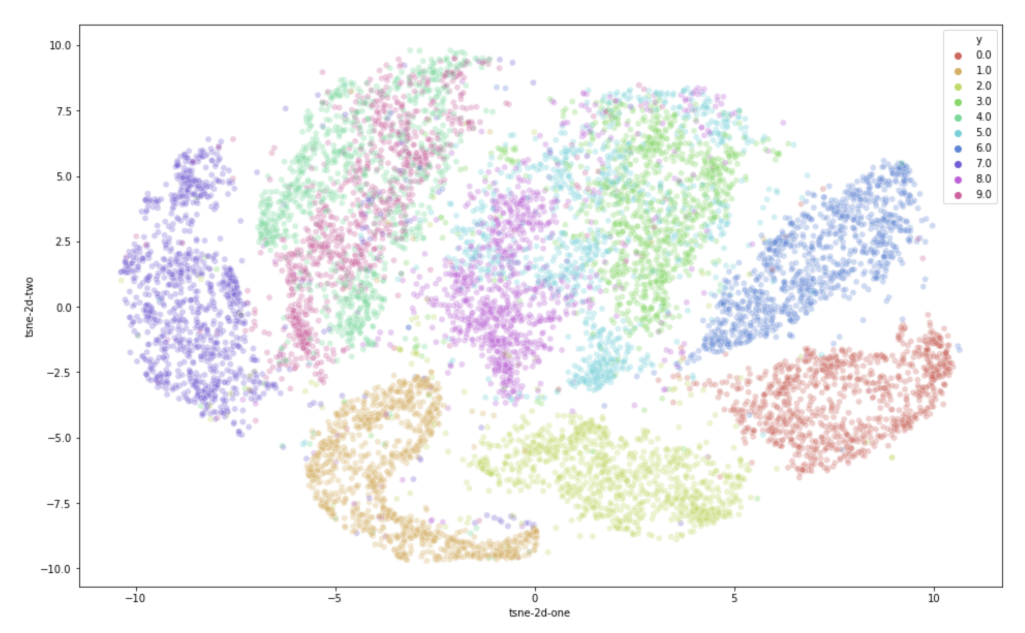


Figure - T-Sne exemple

Les graphes T-sne devrait donner un résultat tel que la figure ci-dessus[[10]](#footnote-10). On constate clairement la démarcation entre les différentes zones.

Cette technique a été utilisée dans mon projet pour des modèles de machines learning et de deep learning.

#### Glove

Glove est un algorithme de machine learning non supervisé permettant de représenter les phrases dans l’espace. La distance entre les mots est liée à leur similitude sémantique. Cet algorithme est Open-source er a été développé par l’université de Stanford[[11]](#footnote-11).

Sur le site de Glove[[12]](#footnote-12), il est possible de récupérer un fichier ou un vecteur représentant un mot a été précalculé à l’avance. Ces vecteurs ont été précalculés sur une très grande quantité de données. Le fichier que j’ai choisi représente un mot avec un vecteur à 100 composantes. Le fichier se présente sous la forme d’un fichier csv ou la première colonne indique le mot et les 100 suivantes indiquent les composants du vecteur.

Pour représenter un mot dans l’espace, il suffit alors de regarder si ce mot est présent dans le dictionnaire. Si le mot est présent, on récupère le vecteur sinon on lui attribue un vecteur avec 100 composantes à 0.

Cette méthode a été utilisée dans mon projet pour les modèles de deep learning.

### Padding

Pour la technique du padding, il suffit d’attribuer à chacun des mots du jeu de données un indice. Chaque mot est alors représenté dans une phrase par son indice. Comme il faut que toutes les données envoyées au modèle soient de la même taille, il faut calculer la taille de la plus longue des phrases du jeu de données. Toutes les phrases envoyées au modèle devront faire cette taille. Les mots manquants sont comblés pas l’indice 0.

Par exemple, si la phrase : [ceci(34) est(2) une(4) phrase(7)] où les chiffres en parenthèse représente l’indice du mot dans le jeu de données et que la taille de la plus grande phrase du jeu de données est de 8, cette phrase deviendra alors : [34, 2, 4, 7, 0, 0, 0, 0]. C’est sous cette forme que la phrase sera envoyée au modèle.

Cette technique a été utilisée dans mon projet pour les modèles de deep learning.

## Modèle de machine learning

Une multitude de modèles de machine Learning ont été testés pour ce projet. En effet, dans sklearn, le changement d’un modèle pour un autre se fait sans difficulté.

### Modèle de régression

Voici une liste des modèles utilisés :

* La régression linéaire (LinearRegression)
* La machine à vecteur de support (SVR)
* La machine à vecteur de support linéaire (LinearSVR)
* La forêt aléatoire (RandomForestRegressor)

La régression linéaire a été choisie, car c’est un modèle de machine learning basique pour les tâches de régression. La régression linéaire est surtout utilisée pour s’assurer que les autres modèles ont de meilleurs résultats que ce modèle qui est basique. Les 2 autres modèles ont été sélectionnés, car ce sont 2 modèles puissants pour les tâches de régression. La machine à vecteur de support linéaire est utilisée dans le cadre de jeu de données de grande taille comme par exemple celui des tweets financiers. En effet le SVR classique à une complexité algorithmique qui croît fortement pour plus de 10'000 observations[[13]](#footnote-13).

Le modèle de machine learning est ajouté à la suite du pipeline. Les différents modèles ont alors été entraînés sur le jeu d’entraînement.

### Modèle de classification

* La classification naïve bayésienne (MultinominalNB)
* La régression logistique (LogisticRegression)
* La machine à vecteur de support (SVC)
* La machine à vecteur de support linéaire (LinearSVC)
* La forêt d’arbre de régression (DecisonTreeClassifier)

La classification naïve bayésienne et la régression logistique sont comme pour les tâches de régression des modèles basiques qui permet de s’assure que les autres modèles apprennent bien à partir des données. La machine à vecteur de support et la forêt d’arbre de décision sont comme pour les tâches de régression, de puissant modèle. Le LinearSVC permet une nouvelle fois comme pour la régression de pouvoir traiter des jeux de données de grande taille.

Il est également possible avec des modèles de régression d’obtenir la précision des données prédites comme nous le donnerait un classificateur. Pour effectuer cette étape, les prédictions du modèle devront être placées dans la classe positive ou négative selon un certain seuil. Cette manière de faire peut donner de mauvais résultat si beaucoup de prédictions ont une polarité proche de 0.

## Modèle de deep learning

Voici les différents types de réseau de neurones que j’ai testé pour mon projet. Les architectures sont détaillées dans la section qui parle des résultats de ces modèles.

* LSTM
* Recurrent Neural Network
* GRU
* Convolutional Neural Network

## Évaluation des modèles de machines learning

Les scores dont je parle dans les chapitres suivants peuvent changer de quelques pour cent entre l’exécution du notebook et le rapport. Ceci vient du fait qu’il y a tellement de petits paramètres pour un modèle de machine learning qu’un petit changement peut engendrer un score légèrement différent. Cependant l’ordre de l’erreur reste la même. Les étapes de conception de ces modèles peuvent se trouver dans le notebook « *HeadlinesMachineLearningModel.ipynb* ».

Les captures suivantes montrent à chaque fois le modèle ayant le meilleur score pour cette étape.

### Titres d’articles financiers

Pour chacune des étapes, j’ai prétraité les phrases à l’aide de Spacy. J’ai en effet remarqué que le modèle donnait de meilleurs scores qu’avec le prétraitement de Gensim. Les étapes de prétraitement effectué sont :

* Tokenisation
* Suppression des « stop-word »
* Suppression de la ponctuation
* Lemmatisation du mot

Voici les différents scores que j’obtiens avec ce jeu de données. Pour chacune des étapes, c’est le modèle avec le meilleur score qui a été conservé.

#### Modèle avec le Tf-Idf

Voici le score du modèle utilisant le Tf-Idf et le prétraitement effectue avec Spacy. On voit que par rapport à VADER et Textblob, le résultat est bien meilleur. On remarque que l’on atteint une exactitude de 80%. Le modèle qui obtient les meilleurs résultats est une forêt aléatoire.

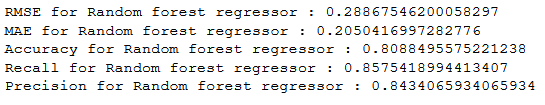


Figure - Score modèle Tf-Idf - Headlines

##### Recherche par quadrillage

Pour prouver que le modèle ne surévalue pas trop et pour optimiser les hyperparamètres j’ai entraîné le modèle avec une recherche aléatoire (RandomizedSearchCV dans sklearn). Cette fois si le modèle est entraîné en 5 blocs (fold) différents. Le jeu d’entraînement est découpé aléatoirement en 5 blocs distincts, puis l’entraînement et l’évaluation s’effectuent en 5 passes successives. Un bloc est à chaque phase gardé pour l’évaluation et les 4 autres pour l’entraînement[[14]](#footnote-14). Ce processus permet de s’assurer que le modèle ne surévalue pas trop.

Cette étape est alors réalisée plusieurs fois en changeant les hyperparamètres du modèle (pour ce modèle j’ai choisi 50 itérations). J’ai choisi d’effectuer la recherche aléatoire avec la forêt aléatoire, car c’est ce modèle qui donnait les meilleurs résultats à l’étape précédente.

Voici les hyperparamètres utilisés :

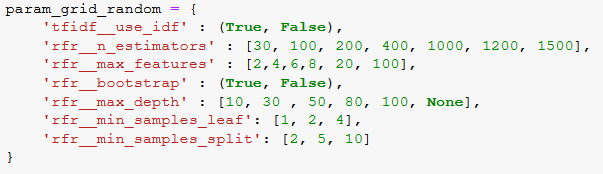
La clé du dictionnaire correspond aux hyperparamètres à tester et la valeur du dictionnaire correspond aux valeurs à tester pour le modèle.

Figure - Recherche aléatoire paramètre

La recherche aléatoire va entraîner le modèle avec des paramètres aléatoires pour chaque itération contrairement à la recherche par quadrillage qui va tester toutes les combinaisons possibles d’hyperparamètres.

Après avoir entraîné le modèle avec la recherche aléatoire, on constate que les résultats se sont améliorés. Le modèle atteint une exactitude de 83 %. De plus comme le modèle a été entraîné par validation croisée, cela augmente les chances qu’il ne surévalue pas.

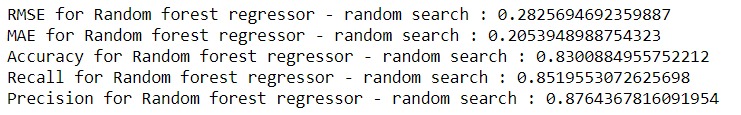


Figure - Recherche aléatoire score

Voici les différents hyperparamètres de ce modèle :

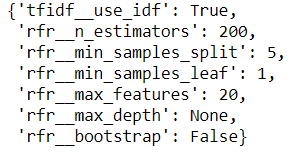


Figure - Recherche aléatoire meilleurs paramètres

#### Modèle avec Doc2vec de Gensim

Pour ce modèle, j’ai utilisé la classe Doc2Vec de Gensim pour représenter mes données dans l’espace. Le prétraitement du texte a été effectué avec Spacy. Le modèle qui obtient les meilleurs résultats est de nouveau la forêt aléatoire. Cependant, on voit cette fois-ci que le modèle obtient de moins bons résultats que précédemment. L’exactitude est moins bonne d’environ 10% et le compromis précison/rappel est moins bon. De plus la RMSE et la MAE ont augmenté.

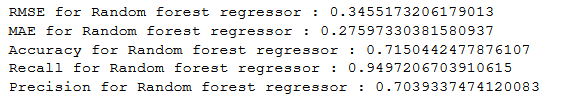


Figure - Scores modèle doc2vec - Headlines

J’ai également entraîné la forêt aléatoire toujours avec les mêmes données en utilisant une validation croisée en 5 phases pour prouver que le modèle ne surévalue pas. 

Figure - Scores modèle doc2vec - validation croisée - Headlines

Voici les scores de la validation croisée. Chaque élément de la liste correspond au score d’une des phases de la validation croisée. On constate que la MAE au même ordre de grandeur que celle au point précédent. Cela augmente les chances que le modèle ne surévalue pas.

#### Modèle avec le Word2Vec financier

Pour ce modèle j’ai utilisé le prétraitement du texte de Spacy. Pour la représentation de la phrase dans l’espace j’ai utilisé la moyenne des vecteurs de mots multipliés par le score Tf-Idf. J’ai utilisé cette représentation, car c’est celle qui sépare le mieux les phrases positives des négatives tout en gardant un bon score par rapport aux autres représentations spatiales présentées au chapitre « Word2Vec financiers ». Le meilleur modèle est encore une fois une forêt aléatoire. L’utilisation de la foret aléatoire comporte l’avantage supplémentaire qu’il diminue la chance de surajustement comme le modèle est entraîné sur plusieurs arbres de décision.

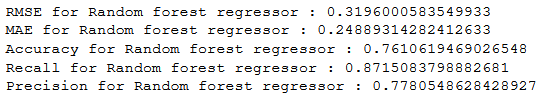


Figure - Scores modèle word2vec - Headlines

Les autres modèles comme le SVC donnent aussi de bons résultats, mais à peine moins bon que la forêt aléatoire. C’est pour cette raison que je garde uniquement la forêt aléatoire. Pour ce cas, par exemple, le SVC donnait une précision de 75,9% contre 76,1 pour la forêt aléatoire.

Après avoir entraîné ce modèle avec la validation croisée de 5 phases j’obtiens les résultats suivants :



Figure - Scores modèles word2vec - validation croisée - Headlines

On constate que la MAE n’a pas augmenté par rapport au résultat précédent. Cela diminue donc les chances que le modèle surévalue.

##### Modèles avec différentes représentation vectorielle

Voici les résultats obtenus si l’on entraîne une forêt aléatoire, mais en utilisant une autre représentation dans l’espace des données :

* Pour la concaténation des vecteurs :
  + Exactitude : 75,4 %
  + MAE : 0.25
* Pour la somme multipliée par le score Tf-Idf :
  + Exactitude : 77,1 %
  + MAE : 0.24

On constate que ces 2 modèles ont des scores semblables à celui obtenu avec la moyenne des vecteurs. Cependant ces 2 modèles ont tendance à moins effectuer la démarcation entre les phrases positives et négatives même si le résultat est minime. De plus, pour la concaténation de vecteur le vecteur en entrée contient dans mon cas 2500 composantes. Plus il y a de composantes plus le modèle est complexe et a de chance de surévaluer.

C’est pour ces différentes raisons que mon choix s’est plutôt portées sur la moyenne des vecteurs multiplié par le score Tf-Idf.

#### Évaluation des modèles sur les titres d’articles financiers annotés par les experts

J’ai sélectionné les 3 modèles suivants pour effectuer des prédictions sur les headlines financières annotées :

* Forêt aléatoire avec Tf-Idf
* Arbre de décision (classificateur) avec Tf-Idf
* Forêt aléatoire avec la moyenne des vecteurs de mot

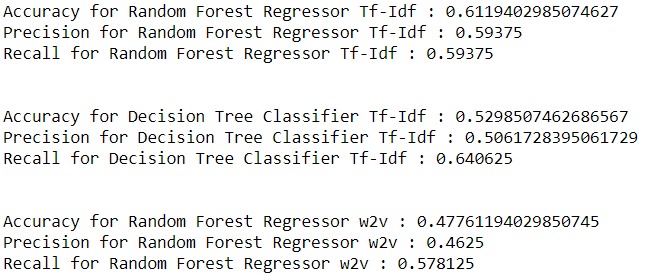
Avec la capture ci-dessus on voit que les scores sont très mauvais.

Figure - Scores des modèles avec les Headlines financières annotés par les experts

On constate que les modèles ont une précision d’environ 60%. Cela veut dire qu’il effectue des prédictions de manière quasiment aléatoire. Ces modèles obtiennent environ les mêmes scores que Vader et Textblob. Les différentes raisons de l’obtention de ces scores sont expliquées en détail dans la conclusion sur les solutions développées.

### Tweets financiers

Voici les différents scores que j’obtiens avec ce jeu de données. Pour chacune des étapes, c’est le modèle avec le meilleur score qui a été conservé. Les différentes étapes de conception de ces modèles se trouvent dans le notebook « *TweetsMachineLearningModel.ipynb* » :

Étant donné, le grand nombre de données à ma disposition plus de …, il m’a fallu réduire la taille du jeu de donnée. Les scores suivants ont été obtenus avec 30'000 données. Pour augmenter légèrement les scores et diminuer le risque que le modèle surévalue, il est possible d’augmenter le nombre de données utilisées. Cependant le temps d’exécution des cellules prendra beaucoup plus de temps.

#### Modèles avec le Tf-Idf

Pour ces modèles j’ai utilisé le prétraitement de Gensim et la représentation Tf-Idf pour les phrases. Le modèle présenté ci-dessous est un SVR. On constate que le modèle à de très bons résultats.

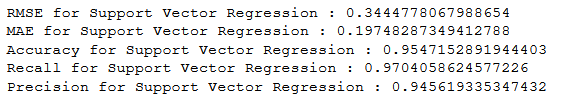


Figure - Scores modèle Tf-Idf - Tweets

Le modèle obtient une exactitude de 95% et un bon taux rappel/précision. La MAE et la RMSE ont également de bons résultats.

Si l’on entraîne le modèle avec la validation croisée, on remarque que la MAE est semblable au point précédent. Le modèle n’a donc pas l’air de surajuster. De plus comme le jeu de données contient pas mal de données, les chances de surajuster diminue. J’ai ici entraîné les modèles en validation croisée avec 15 phases comme le jeu de données est de plus grande taille.



Figure - Scores modèle Tf-Idf - validation croisée - Tweets

Pour conclure, on peut dire que ce modèle semble intéressant.

#### Modèle avec Doc2vec de Gensim

Les phrases de ce modèle ont été prétraitées avec Gensim et elles sont représentées dans l’espace à l’aide de la classe Doc2Vec de Gensim. Le meilleur modèle pour cette étape est une régression linéaire.

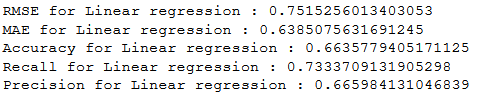


Figure - Scores modèle doc2vec - Tweets

On constate avec ce modèle que les performances du modèle se sont considérablement dégradées. Quand on regarde la RMSE et la MAE on constate que les prédictions sont très éloignées de la réalité. Le modèle à de la peine à généralisé. Les autres modèles développés pour cette étape ont aussi de mauvaises performances.

#### Modèle avec le Word2Vec financier

Les phrases de ce modèle ont été prétraitées avec Gensim et la représentation dans l’espace s’effectue en faisant la moyenne des vecteurs de mot multiplié par le score Tf-Idf. J’ai choisi cette représentation dans l’espace pour les mêmes raisons que celle évoquée avec le jeu de données des titres d’articles financiers. Voici les résultats obtenus :

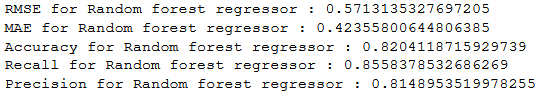


Figure - Scores modèle word2vec - Tweets

On constate que le modèle à une bonne exactitude, rappel et précisons. Mais, la RMSE et MAE sont quand même élevées. Cela veut dire que le modèle fait pas mal d’erreurs, mais arrive quand même à trouver la bonne classe pour les phrases.

Si on entraîne le modèle avec une validation croisée de 10 phases, on constate que la MAE est semblable au point précédent donc le modèle n’a pas l’air de surajuster.



Figure - Scores modèles word2vec - validation croisée - Tweets

#### Classificateur avec Tf-Idf

Pour cette étape. J’ai enlevé du jeu de données les phrases qui avaient un sentiment entre -0.3 et 0.3. J’ai également converti le sentiment de la phrase en des valeurs discrètes 0 ou 1. Pour cette étape les phrases ont été prétraitées avec Gensim. Le meilleur modèle est ici un SVC. Les autres modèles testés ont également de très bons résultats.

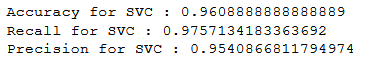


Figure - Scores modèle classificateur - Tweets

On constate que l’exactitude atteint les 96% et que le compromis précision rappel est très bon. Si l’on entraîne le modèle avec une validation croisée en 10 phases on constate que le modèle n’a pas l’air de surajuster. L’exactitude ne change pas par rapport à avant.



Figure - Scores modèle classification - validation croisée - Tweets

#### Calcul du score sur les titres d’articles financiers

Ce modèle a été entraîné uniquement avec des tweets. Cependant ce qui nous intéresse dans ce projet ce sont les titres d’articles financiers. Il faut donc tester les modèles précédemment créés sur le jeu de données des titres d'articles financiers. J’ai alors utilisé tous les titres d’articles financiers que j’ai à ma disposition. J’ai alors sélectionné les modèles suivants pour effectuer des prédictions sur les titres d’articles :

* SVC avec prétraitement Gensim et Tf-Idf

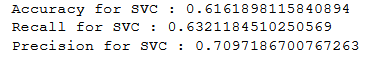


Figure - Scores modèle SVC - Headlines

* SVR avec prétraitement Gensim et Tf-Idf

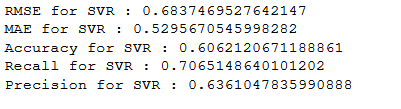


Figure - Scores modèle SVR - Headlines

* Foret aléatoire avec prétraitement Gensim et Moyenne des vecteurs de mot

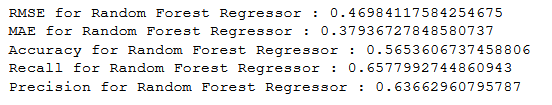


Figure - Scores modèle RFR - Headlines

On constate dans les trois cas que les performances se sont dégradées considérablement. La précision est à peine en dessus de 50%. Le compromis rappel et précision est bas. La MAE et RMSE sont également élevées.

On remarque donc que les modèles entraînés avec les tweets sont très bons sur les tweets, mais mauvais sur les titres d’articles financiers. On peut donc dire que les modèles surajustent.

Cette solution n’est donc pas bonne et est à abandonner.

### Tweets quelconques

Étant donné que les tweets financiers extraits de l’ensemble des tweets à ma disposition ne sont pas de qualité, je ne développe pas les résultats obtenus dans le rapport. J’ai tout de même créé un notebook jupyter du nom de « TweetsMachineLearningModel.ipynb » qui retrace les différentes des étapes pour la création de modèles avec ce jeu de données. Les étapes du notebook et les modèles développés sont semblables aux étapes expliquées précédemment pour les autres jeux de données. Les résultats obtenus ne sont pas et ne dépasse pas les 65 %. De plus comme les tweets sont très différents des titres d’articles financiers, le modèle aurait donné encore de plus mauvais résultats pour des prédictions à partir de tweet financier.

## Évaluation des modèles de deep learning

Les modèles de deep learning ont été développés à partir des titres d’articles financiers. Les différentes étapes de conception de ces modèles sont détaillées dans le notebook « *HeadlinesDeepLearningModel.ipynb* ».

### LSTM avec padding

Pour ce modèle j’ai représenté les phrases à l’aide de la technique du padding.

Ce modèle contient comme première couche, une couche Embedding qui prend en paramètre une dimension d’entrée qui correspond à la taille du vocabulaire et la dimension de sortie que j’ai fixée à 32. La seconde couche est une couche LSTM[[15]](#footnote-15) de taille 32. La dernière couche du modèle est une couche de type Dense avec une sortie de taille 1 et une fonction d’activation sigmoid.

L’architecture en résumé :

* Emedding(entrée : taille du vocabulaire, sortie : 32)
* LSTM(entrée : 32, sortie : 32)
* Dense(entrée : 32, sortie : 1)

J’ai entraîné ce modèle sur 20 epoch avec une taille d’échantillon de 64. Voici les résultats de ce modèle avec un graphe de l’exactitude en fonction des epochs et un graphe de la perte en fonction des epochs.

Figure - Scores modèle LSTM

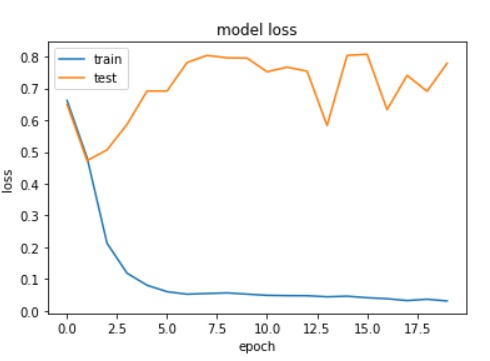
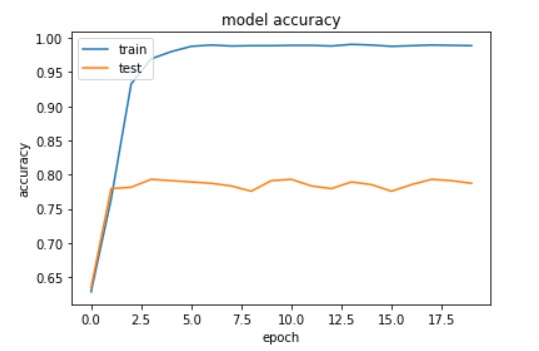


Figure - Graphe LSTM accuracy

Figure - Graphe LSTM loss

On constate que les performances à l’entraînement sont excellentes, mais dès qu’on passe sur le jeu de tests les performances deviennent médiocres. En regardant le graphe de droite, on constate que la perte ne diminue pas voir pire qu’elle augmente. Le modèle n’arrive pas à apprendre à partir du jeu d’entraînement. Autrement dit le modèle surajuste beaucoup. Même en diminuant la complexité du modèle en réduisant la dimension de sortie de la couche Emedding et la dimension de la couche LSTM le modèle surajuste toujours autant et les performances sont toujours médiocres.

Il y a trop peu de données pour entraîner un réseau de neurones et les phrases à ma disposition sont trop courtes. Ce type de réseau de neurones ne donne donc pas de bonne performance.

### GRU avec padding

Ce modèle utilise également la technique du padding pour représenter une phrase. La configuration de ce modèle est la même que celle du précédent à l’exception que la dimension de sortie de la couche Embedding de 16, la couche LSTM est remplacée par une couche GRU de taille 16.

L’architecture en résumé :

* Emedding(entrée : taille du vocabulaire, sortie : 16)
* GRU(entrée : 16, sortie : 16)
* Dense(entrée : 16, sortie 1)

J’ai de nouveau entraîné ce modèle avec 20 epoch et une taille d’échantillon de 64. Voici les résultats et les graphes d’exactitude et de perte.

Figure - Scores modèle GRU

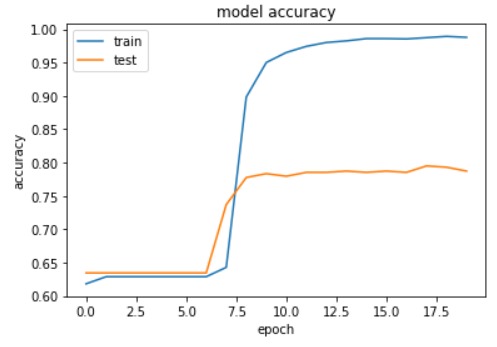
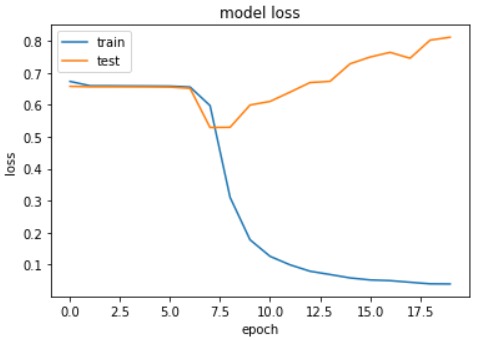


Figure - Graphe GRU accuracy

Figure - Graphe GRU loss

On constate que comme le modèle précédent, ce modèle surajuste beaucoup. La perte augmente au lieu de diminuer et l’exactitude n’augmente.

Comme pour le cas d’avant il faudrait plus de phrases à disposition.

### CNN avec Word2Vec de Glove

Pour ce modèle j’ai utilisé le Word2Vec fourni par Glove. J’ai alors créé un dictionnaire de la taille du jeu de données. La clé de ce dictionnaire est l’index du mot dans le vocabulaire et la valeur est le vecteur correspondant au mot. Ce vecteur provient du fichier fourni par Glove. Si un mot ne se trouve pas dans fichier il est remplacé par un vecteur avec toutes ses composantes à 0.

La taille finale du dictionnaire est la taille totale du vocabulaire et j’ai fixé la taille des vecteurs à 32.

Les phrases sont comme pour les 2 étapes précédentes représentées à l’aide de la technique du padding.

L’architecture du modèle contient une première couche de type Embedding. Je dois donner en paramètre à cette couche le dictionnaire de vecteur créé précédemment. C’est ce qui permettra au modèle de lier les indices des mots à leurs vecteurs. La seconde couche est de type GlobalMaxPool1d. La suivante est une couche de type Dense avec une dimension 10 de sortie et une fonction d’activation relu. La dernière couche est une couche Dense avec une dimension de sortie de 1 et une fonction d’activation Sigmoid.

L’architecture en résumé :

* Emedding(entrée : taille du vocabulaire, sortie : 32)
* GlobalMaxPool1d(entrée : 32, sortie : 32)
* Dense(entrée : 32, sortie : 10)
* Dense(entrée : 10, sortie 1)

J’ai entraîné ce modèle sur 20 epochs avec une taille d’échantillon de 64. Voici les résultats et graphes de ce modèles.



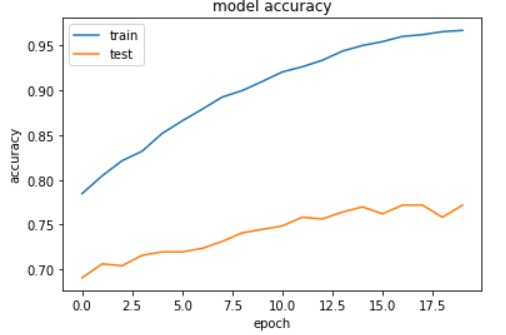
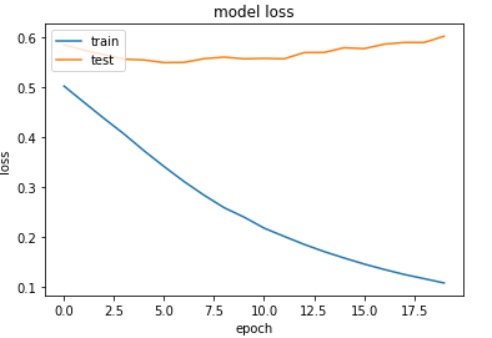
Figure - Scores modèle CNN

Figure - Graphe CNN loss

Figure - Graphe CNN accuracy

On constate encore une fois que le modèle surajute beaucoup trop et que la perte diminue au lieu d’augmenter. Le modèle n’arrive rien à apprendre du jeu d’entraînement.

### RNN avec Word2vec financier

Les étapes pour créer ce modèle se trouvent dans le notebook « *HeadlinesRNNModel.ipynb* ».

Pour ce modèle j’ai représenté comme pour les précédents les phrases avec la technique du padding.

J’ai ensuite utilisé le Word2Vec financier entraîné sur plusieurs millions de mots dans le domaine de la finance.

J’ai alors créé une matrice de la taille totale du vocabulaire multiplié par la taille des vecteurs que j’ai fixé à 100. Le numéro de la colonne de la matrice représente l’index du mot dans le vocabulaire. Une ligne de la matrice représente les composantes du vecteur pour le mot donné. Si le mot n’existe pas dans le Word2Vec il est remplacé par un vecteur avec toutes ses composantes à 0.

L’architecture de ce modèle est composée d’une première couche Embedding qui a une taille d’entrée de la dimension du vocabulaire et une taille de sortie de la taille d’un vecteur représentant un mot c’est-à-dire 100. Il faut également passer en paramètre à cette couche la matrice créée précédemment. La seconde couche est une couche SimpleRNN de taille 16. La dernière couche est une couche Dense avec une taille de sortie de 1 et une fonction d’activation Sigmoid.

L’architecture en résumé :

* Emedding(entrée : taille du vocabulaire, sortie : 100)
* SimpleRNN(entrée : 100, sortie 16)
* Dense(entrée : 16, sortie 1)

Voici les résultats et les graphes obtenus avec ce modèle.



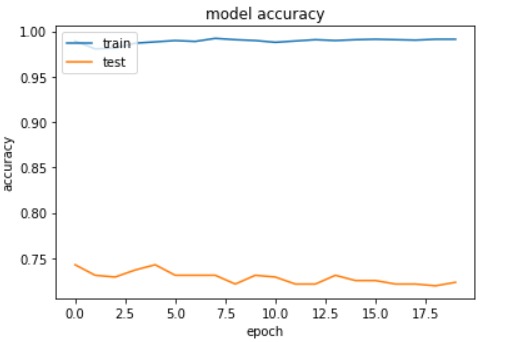
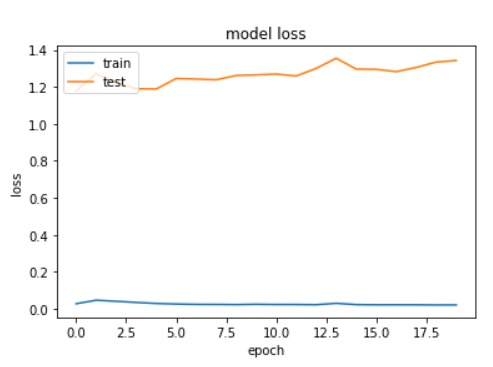
Figure - Scores modèle RNN

Figure - Graphe RNN loss

Figure - Graphe RNN accuracy

On constate que le fait que les vecteurs de mot ont 100 composantes fait que le modèle surajuste énormément. La perte est très élevée et l’exactitude de monte pas. Pour avoir de meilleur résultat, il aurait fallu garder seulement par exemple les 25 premières composantes des vecteurs de mot.

Ce modèle comme tous les autres modèles de deep learning présenté précédemment ne sont pas utilisable car ils surajustent beaucoup trop. Les réseaux de neurones ont besoin de beaucoup de donnée pour donner de bon résultat. Le jeu de données mis à ma disposition ne contient clairement pas assez de phrases pour qu’un réseau de neurones donne des résultats satisfaisants.

## Conclusion des solutions développées

Pour ce qui est des solutions de deep learning on constate que les résultats ne sont vraiment pas bons. Les performances sur le jeu d’entraînement sont excellentes, mais celles sur le jeu de test sont très mauvaises. Les différents modèles surajute beaucoup trop, la perte ne diminue pas d’une epoch à l’autre (au contraire elle monte) et l’exactitude n’augmente pas. Le problème vient du fait que les modèles de deep learning ont besoin d’une grande quantité de données pour donner de bons résultats. Les 2800 titres d’articles financiers à ma disposition ne sont clairement pas assez suffisants pour donner de bons résultats.

Pour ce qui est des tweets financiers, on constate que les modèles donnent de très bons résultats quand les données de tests sont également des tweets financiers. Dès que l’on teste le modèle avec les titres d’articles financiers on voit que les performances diminuent beaucoup. On constate donc que la structure des titres d’articles financiers et des tweets financiers est trop différente l’une de l’autre. Le modèle n’arrive pas bien à généraliser et donne de mauvais scores.

Les modèles de machines Learning avec les titres d’articles financiers donnent à première vue de meilleurs résultats que les solutions précédentes. De plus, étant donné que les modèles sont des forêts aléatoires entraînées en validation croisée cela devrait diminuer les chances que les modèles surajute. Cependant à la vue des résultats sur les données de validation on constate que les scores de ces modèles ne sont pas bons et qu’ils obtiennent des scores similaires à Vader et Textblob. Plusieurs raisons peuvent expliquer ces scores.

Premièrement comme les graphes T-sne l’ont montré précédemment il n’y a pas de démarcation précise entre les phrases positives et négatives. Cela signifie que les différents modèles auront de la peine à apprendre quelque chose et généraliser.

Deuxièmement, comme expliquer précédemment les titres d’articles financiers devant me servir comme donné de validation ont été annoté par rapport à la montée ou la descente du dollar par rapport à l’euro. Mes données initiales ont quant à elle été annotées à partir du contexte des mots dans la phrase. Par exemple une phrase avec plusieurs mots qui ont un sens négatif, mais qui influence le dollar positivement. Mon modèle la classera comme phrase négative, car il se basera sur les mots. Mais la phrase aura été annotée positivement, car elle fait monter le cours du dollar. Mon modèle effectuera donc une erreur.

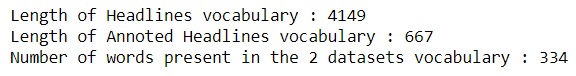
Dernièrement, comme le montre la figure ci-dessous. On constate que sur les 4149 présents du vocabulaire du jeu de données initial seul 334 sont présents sur les 667 du jeu de données. Les modèles se trouveront confronter à de nombreux mots qu'il ne connaisse pas lors de leurs prédictions. Les scores Tf-Idf des mots, certains vecteurs représentant un mot ne pourront pas être calculés et seront alors représentés par leur valeur par défaut (0 pour le Tf-Idf et un vecteur rempli de 0 pour le vecteur de mot). Ceci peut expliquer les mauvais scores obtenus. Vader et Textblob obtiennent sûrement de bons résultats, car ils ont été entraînés sur beaucoup de mots. On constate donc que la couverture du jeu de données n’est pas bonne et qu’il faudrait donc plus de données pour de meilleurs résultats.

Figure - Couverture du vocabulaire

Pour conclure, on peut dire qu’avec les données initiales qui m’ont été fournies il est difficile d’obtenir un meilleur score que Vader et Textblob pour prédire des titres d’articles financiers annotés sur la montée ou la descente du dollar. Pour augmenter les scores des modèles, il faudrait disposer de beaucoup plus de titres d’articles financiers annotés de cette manière. Les modèles de machines learning et de deep learning verrait alors leurs scores s’améliorer et surpasseraient les résultats de Textblob et Vader.

# Approfondissement possible

## Utilisation de Bert

Bert[[16]](#footnote-16) est un modèle spécialisé dans les tâches d’analyse de langage (NLP). C’est un modèle préentraîné qui a été développé par Google. Ce modèle donne de très bons résultats pour l’analyse de sentiment. Il serait incessant de développer un modèle en utilisant cette technologie pour visualiser les scores qu’il est possible d’atteindre.

## Agrandissement du jeu de données

EDA[[17]](#footnote-17) (Easy Data Augmentation) est un ensemble de technique permettant d’augmenter la taille d’un jeu de données textuelles. Les techniques proposé ici sont :

* Remplacement d’un nombre aléatoire de mots dans la phrase par leur synonyme.
* Insertion d’un nombre aléatoire de synonymes dans la phrase à partir des mots de cette phrase.
* Inversion de deux mots dans la phrase plusieurs fois.
* Suppression de mot aléatoire dans la phrase.

Cette technique permet de créer de nouvelle phrase et donc d’augmenter la taille du jeu de données. Cette solution donne apparemment de bons résultats sur les petits jeux de données comme celui que j’ai à disposition.

Selon la quantité de phrases générée, il pourrait être intéressé d’exécuter les modèles de machine learning et de deep learning sur ces nouvelles données pour voir si les scores des modèles augmentent.

# Conclusion

Dans le cadre de ma formation, c’est la première fois que j’avais à réaliser un projet de machine learning. La réalisation de ce type de projet est très différente d’un projet classique dans l’informatique. J’ai également dû passer pas mal de temps à faire des recherches sur les fondamentaux du machine learning étant donné que je suis débutant de domaine. Une grande partie du temps du projet est consacrée à la recherche et aux tests de différentes solutions. Étant donné mon projet de recherche il m’a fallu m’adapter au fait de recevoir plusieurs jeux de données au cours du projet. J’ai dû recommencer le processus d’élaboration d’un modèle plusieurs fois.

J’ai également été handicapé par le fait de recevoir le jeu de données permettant de valider mon travail l’avant-dernière semaine du projet. Je me suis donc aperçu tardivement que mon modèle ne donnait pas de bon résultat sur les données fourni par les experts financiers. Il m’a donc fallu trouver rapidement des réponses à ces résultats.

Pour conclure sur les résultats obtenus, on peut donc constater qu’avec la taille du jeu de données initial qui m’a été fourni je n’obtiens pas de meilleurs résultats que des analyseurs de sentiment entraîné sur tout le vocabulaire de la langue anglaise (Vader et Textblob). Les modèles développés à l’aide des titres d’articles financiers sont efficaces pour faire des prédictions sur le même type de données. Ils n’obtiennent pas de bons résultats à cause du fait que les phrases positives et négatives n’ont pas de séparation claire dans l’espace, les phrases annotées par les experts n’ont pas été étiquetées de la même manière que mon jeu de donné initial et que la taille de mon jeu de donné initial est trop petit.

Pour augmenter les scores des modèles, il faudrait les entraîner sur un jeu de données de grande taille annotée de la même manière que les titres d’articles financiers labélisés par les experts.

Malgré les résultats obtenus, je pense avoir acquis les différents objectifs pour ce projet en tenant compte du fait que je n’avais jamais fait de machine learning. Ce projet m’a permis de faire une très bonne introduction au domaine du machine learning et m’as permis d’acquérir plein de nouvelles compétences dans le domaine.

# Bibliographie

## Sites web

* <https://medium.com/>
* <https://www.wikipedia.org/>
* <https://stackoverflow.com/>
* <https://scikit-learn.org/stable/user_guide.html>
* <https://keras.io/>
* <https://www.kaggle.com/>
* <https://towardsdatascience.com/>
* <https://stackabuse.com/>
* <https://spacy.io/api/doc>
* <https://fr.quora.com/>
* <https://radimrehurek.com/gensim/>
* <https://realpython.com/>
* <https://stackexchange.com/>
* <https://pandas.pydata.org/>
* <https://www.nltk.org/>
* <https://textblob.readthedocs.io/en/dev/>
* <https://nlp.stanford.edu/projects/glove/>
* <https://github.com/>

## Livres

* Machine learning avec Scikit-Learn, Aurélien, Géron, Paris, Dunod
* Deep Learning avec TensorFlow, Aurélien, Géron, Paris, Dunod

# Annexes

## Dépendances

* **python : 3.7.4**
* jupyter notebook : 6.0.1
* sklearn : 0.21.3
* matplotlib : 3.1.1
* pandas : 0.25.1
* numpy : 1.17.4
* csv : 1.0.0
* vaderSentiment : 3.2.1
* textblob : 0.15.3
* gensim : 3.8.1
* joblib : 0.13.2
* seaborn : 0.9.0
* spacy : 2.2.1
* re : 2.2.1
* en\_core\_web\_sm : 2.2.0
* tensorflow : 2.0.0
* keras : 2.3.1
* pickle : 4.0.0

## Journal de bord

1. <https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-1/src/master/> [↑](#footnote-ref-1)
2. <https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-2/src/master/> [↑](#footnote-ref-2)
3. <https://spacy.io/> [↑](#footnote-ref-3)
4. <https://tedboy.github.io/nlps/generated/generated/gensim.utils.simple_preprocess.html> [↑](#footnote-ref-4)
5. <https://docs.scipy.org/doc/scipy/reference/sparse.html> [↑](#footnote-ref-5)
6. <https://radimrehurek.com/gensim/> [↑](#footnote-ref-6)
7. <https://nlp.stanford.edu/projects/glove/> [↑](#footnote-ref-7)
8. <https://radimrehurek.com/gensim/sklearn_api/d2vmodel.html> [↑](#footnote-ref-8)
9. <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html> [↑](#footnote-ref-9)
10. <https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b> [↑](#footnote-ref-10)
11. <https://en.wikipedia.org/wiki/GloVe_(machine_learning)> [↑](#footnote-ref-11)
12. <https://nlp.stanford.edu/projects/glove/> [↑](#footnote-ref-12)
13. <https://stackoverflow.com/questions/31681373/making-svm-run-faster-in-python> [↑](#footnote-ref-13)
14. Explication de la validation croisée dans le livre « Machine learning avec Scikit-Learn, Aurélien, Géron, Paris, Dunod », page 68. [↑](#footnote-ref-14)
15. <https://keras.io/layers/recurrent/#lstm> [↑](#footnote-ref-15)
16. <https://en.wikipedia.org/wiki/BERT_(language_model)> [↑](#footnote-ref-16)
17. <https://github.com/jasonwei20/eda_nlp> [↑](#footnote-ref-17)