Unsupervised Sentiment Analysis of CIH Bank Reviews Moroccan Dialect

Réalisé par :

Belfadla Fatima Ezzahra
Belfadla.fati@gmail.com

Kawtar Zinoune
kawtarzinoune45@gmail.com

Encadré par :

- Pr. Abderrahim BENI-HSSANE
- Mr. Mohammed KASRI

Résumé

Les mondes virtuels tels que les sites de réseaux sociaux, les blogs et le contenu des communautés deviennent extrêmement l'une des sources les plus puissantes d'actualités, de marchés, d'industries, etc. Ces mondes virtuels peuvent être utilisés à de nombreux égards, car ce sont de riches plateformes pleines de commentaires, d'émotions, de réflexions et des avis. L'objectif principal de ce travail est de regrouper les avis des clients en arabe marocaine sur la banque CIH pour l'analyse des sentiments en groupes positifs et négatifs. Nous avons utilisé le web scraping pour collecter des avis en arabe associés uniquement aux CIH, à partir du Facebook et YouTube et nous avons obtenu au total 251 avis en arabe. Ensuite, le word2vec a été appliqué pour extraire les caractéristiques numériques de notre dataSet. Une approche d'apprentissage non supervisé a été appliquée, en particulier l'algorithmes K-Means avec une métrique de distance : Cosinus. Nos données de test étiquetées manuellement montrent que l'algorithme K-Means avec la distance cosinus a bien fonctionné lors de l'application de toutes nos prétentions pas.

1. Introduction

Récemment, nous avons assisté à un flux massif de données via des applications intelligentes, des pages Web et des réseaux sociaux. Grâce au développement rapide de technologies de partage d'informations, les gens peuvent partager des opinions, des sentiments et des situations qu'ils ont vécus. Certains sites Web de partage d'informations sont largement utilisés par les demandeurs d'informations. Cela permet aux gens de trouver des banques qualifiées en qui ils peuvent avoir confiance et qui répondent à leurs attentes.

De plus, d'un point de vue économique, les avis des clients sont devenus un facteur important pour choisir une banque, car ils influencent les décisions des nouveaux clients. Les propriétaires des banques peuvent également bénéficier de ces avis, car ils peuvent améliorer leurs services en fonction de l'évaluation des critiques.

Cependant, il est difficile d'utiliser les avis en ligne car ils sont énormes. De plus, les nouveaux clients peuvent avoir du mal à comprendre certaines critiques et déterminer s'ils sont positifs ou négatifs.
Pour pallier ce problème, nous avons proposé une approche basée sur des méthodes d'apprentissage automatique non supervisé pour faire la distinction entre avis positifs et avis négatifs.

En particulier, nous expérimentons des avis en arabe marocaine de la banque CIH qui ont été rassemblés pour les besoins de ce travail.

2. Méthodologie

2.1.DataSet

Les données utilisées ont été collectées à partir de Facebook et YouTube à l'aide d'un crawler. En conséquence, l'ensemble de données contenait 251 avis (fig.1) de CIH Bank appartenant à deux classes : Positifs et Négatifs.

	A	B
1	review	polarity
2	تطبيق جيد، بسيط وفعال لي أنصح بشدة لجميع مالكي الحساب في هذا البنك الجي	1
3	فعال	1
4	أوصي بشدة جميع مالكي الحساب في هذا البنك الجيد جدا.	1
5	بنك جيد جدا!	1
6	علة سيئة!	-1
7	سيئ	1
8	تطبيق رائع، عملي جدا. أصبح كل شيء بسيطا، ويعمل جيدا بالنسبة لي.	1
9	تطبيق رائع	1
10	مريحة للغاية.	1
11	انها تعمل جيدة جدا	1
12	إنه يعمل بشكل جيد بالنسبة لي	1
13	تطبيق جيد جدا أن أستخدمه كل يوم لعرض حساباتي أو إجراء تحويلات	1
14	التطبيق مفيد للغاية	1
15	تطبيق جيد جدا	1
16	عملي للغاية	1
17	5نجوم...	1
18	برافو للتحديث، يفتقد إمكانية إجراء عمليات نقل خارجية. وإمكانية وضع الملاحظ	1
19	البنك الذي يعامل كل طلب بالاحتراف.	1
20	لا يدعم هذا التطبيق حسابات الأعمال، فمن غير مقبول	-1
21	عملي للغاية ممتاز	1
22	برافو Cih برو جميل	1

Fig.1 : Aperçu de la dataSet

Avec :

- -1 représente les avis négatifs
- 1 représente les avis positifs

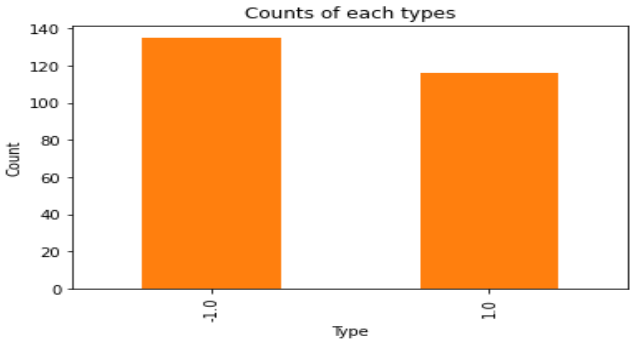

Fig.2 : Distribution des sentiments négatifs et positives

2.2.Prétraitement

Avant de soumettre nos données aux algorithmes d'apprentissage, nous devons les prétraiter en suivant quelques étapes :

1. Suppression des valeurs manquantes.
2. Normalisation et filtrage du texte : Il est nécessaire de nettoyer les avis en supprimant les signes de ponctuation, les caractères spéciaux, les caractères non-arabes, les dates, l'heure, les chiffres, les liens et les diacritiques, etc…
3. Supprimer les Emojis : Les emojis peuvent être considérés comme des données auxiliaires dans la classification des textes en positifs et négatifs, mais dans cette étude, nous nous concentrons uniquement sur l'analyse des textes écrits.

4. Suppression des Stopwords : Les Stopwords sont connus comme mots extrêmement fréquents, tels que (les pronoms, les conjonctions, prépositions et noms). Par conséquent, dans cette étape, nous les avons supprimés.
5. Stemming : Nous n'avons pas utilisé le stemming, car il a changé le contexte des phrases et éliminé presque quelques mots qui sont intéressantes pour la classification

```
data['cleaned_text'] = data.review.apply(clean_text)
data
```

:

	review	polarity	cleaned_text
0	تطبيق جيد، بسيط وفعال لي أنصح بشده لجميع مالكي الحساب في هذا البنك الجيد جدا.	1.0	تطبيق جيد بسيط وفعال انصح بشده لجميع مالكي الحساب البنك الجيد
1	فعال	1.0	فعال
2	أوصي بشده جميع مالكي الحساب في هذا البنك الجيد جدا.	1.0	اوصي بشده مالكي الحساب البنك الجيد
3	إبنك جيد جدا	1.0	بنك جيد
4	إعله سيئه	-1.0	علة سيئه
...	...	...	...
248	وكالات الشركه غير متواجده بكل مكان	-1.0	وكالات الشركه متواجده بكل مكان
249	ينقصه الخدمات	-1.0	ينقصه الخدمات
250	اذا جربتهم كاملين سيرفس ديالهم زوين،	1.0	اذا جربتهم كاملين سيرفس ديالهم زوين
251	ابليكاسيون ديالك كايكون فيها شخص لي مكلف بيك فاش اكون عندك مشكل تتاصل بيه، مكتحتاج تا حاجا كاشي فيها زوين	1.0	ابليكاسيون ديالك كايكون شخص مكلف بيك فاش اكون عندك مشكل تتاصل بيه مكتحتاج حاجا كاشي زوين
252	احسن بنك وشكران cih bank و كتبقا	1.0	كتبقا احسن بنك وشكران

251 rows × 3 columns

Après avoir nettoyé notre corpus, plusieurs autres étapes ont été prises pour préparer les données pour le modèle word2vec. Les principales étapes comprenaient la détection et le remplacement des bigrammes de mots les plus fréquents avec le module Phrases de gensim.
Toutes ces étapes et la plupart des hyperparamètres du modèle Word2Vec que nous avons utilisé étaient basés sur le tutoriel Word2Vec de kaggle [1].

2.3.Modèle

1. Création de modèle word2vec :

   Dans cette partie, on a adopté l'implémentation de gensim de l'algorithme word2vec avec l'architecture CBOW.

   Word2vec est essentiellement une technique d'incorporation de mots utilisée pour convertir les mots de dataSet en vecteurs afin que la machine comprenne. Chaque mot unique dans vos données est affecté à un vecteur et ces vecteurs varient en dimensions en fonction de la longueur du mot.

   Apres avoir implémenter notre modèle et convertir notre dataSet en vecteurs et pour regrouper nos avis en deux classes positif et négatif on va utiliser un algorithme non-supervisé K-Means.

2. Algorithme K-Means :

Il s'agit d'un algorithme de clustering itératif qui vise à partitionner les instances en k clusters dans lesquels chaque observation appartient au cluster avec la moyenne la plus proche.

L'algorithme k-Means est un algorithme d'apprentissage automatique non supervisé qui vise à regrouper les documents en k clusters dans lesquels chaque avis appartient au cluster avec le centroïde le plus proche (McQueen, 1967). Nous avons initialisé le k-Means++ pour accélérer la convergence et exécuter 10 fois avec différents centroïdes et 1000 itérations par exécution.

Après avoir extrait les caractéristiques numériques de notre dataSet, nous avons appelle L'algorithme k-Means avec deux clusters : un pour le positif et l'autre pour le négatif. Nous avons utilisé le cosinus avec l'algorithme k-Means (Huang, 2008).

```
model = KMeans(n_clusters=2, max_iter=10000, random_state=True, n_init=300).fit(X=word_vectors.vectors.astype('double'))
```

L'étape suivante, consiste à attribuer à chaque mot un score de sentiment - une valeur négative ou positive (-1 ou 1) en fonction du groupe auquel ils appartiennent. Pour peser ce score, nous avons la multiplié par la proximité de leur cluster (pour peser à quel point ils sont potentiellement positifs/négatifs). Comme le score que l'algorithme K-Means produit est la distance des deux clusters, pour les peser correctement, nous avons les multipliés par l'inverse du score de proximité (score de sentiment divisé par le score de proximité).

```
words.head(10)
```

	words	vectors	cluster	cluster_value	closeness_score	sentiment_coeff
0	بدك	[-0.0053875335, 0.0024120465, 0.05158014, 0.09...	0	-1	1.001393	-1.001393
1	جيد	[-0.07970168, 0.0899191, -0.0019112608, -0.019...	1	1	1.007716	1.007716
2	التطبيق	[0.08365881, -0.045491967, -0.011122338, 0.010...	1	1	1.020587	1.020587
3	يعمل	[-0.052869197, -0.068392955, -0.079964176, 0.0...	1	1	1.013352	1.013352
4	خدمه	[-0.018804869, -0.05124525, 0.09211358, -0.090...	0	-1	1.009792	-1.009792
5	للغايه	[-0.067311846, 0.0396531, 0.02010653, 0.069740...	1	1	1.008536	1.008536
6	انا	[0.10026633, -0.10072809, -0.0669382, 0.028670...	1	1	1.011398	1.011398
7	تطبيق	[0.0137031805, 0.06540896, 0.0997691, 0.090142...	0	-1	1.014209	-1.014209
8	البنك	[-0.049235113, -0.012671982, 0.03251909, -0.06...	0	-1	1.018231	-1.018231
9	الله	[0.021362126, 0.05691854, -0.020640204, 0.0308...	1	1	1.011678	1.011678

Fig. : échantillon de mots avec coefficients de sentiment pondérés calculés

3. Pondération Tfidf :

L'étape suivante consistait à calculer le score tfidf de chaque mot dans chaque phrase avec le TfidfVectorizer de sklearn. Cette étape a été menée pour voir à quel point chaque mot était unique pour chaque phrase et augmenter le signal positif/négatif associé aux mots qui sont très spécifiques pour une phrase donnée par rapport à l'ensemble du corpus.

```
tfidf = TfidfVectorizer(tokenizer=lambda y: y.split(), norm=None)
tfidf.fit(file_weighting.cleaned_text)
features = pd.Series(tfidf.get_feature_names())
transformed = tfidf.transform(file_weighting.cleaned_text)
```

Enfin, tous les mots de chaque phrase ont été remplacée par leurs scores tfidf, et d'autre part par leurs scores de sentiment pondérés correspondants. Apres on a remplacé les mots dans les phrases par leurs scores tfidf/sentiment associés, pour obtenir 2 vecteurs pour chaque phrase. Le produit scalaire de ces vecteurs à 2 phrases indiquait si le sentiment global était positif ou négatif (si le produit scalaire était positif, le sentiment était positif, et dans le cas contraire négatif).

	sentiment_coeff	tfidf_scores	sentence	sentiment	sentiment_rate	prediction
0	[-1.014209119433242, 1.0077162015668324, 1.002...	[0.23041696120258162, 0.24073829047230175, 0.3...	تطبيق جيد بسيط وفعال انصح بشده لجميع ...مالكي الح	1	0.949568	1
1	[1.0087761004028488]	[1.0]	فعال	1	1.008776	1
2	[1.0153027521154625, 1.0029260599910217, 1.002...	[0.45638891766008133, 0.41401721007667575, 0.4...	اوصي بشده مالكي الحساب البنك الجيد	1	1.730639	1
3	[-1.0013933751414046, 1.0077162015668324]	[0.6767810163327467, 0.7361843898994425]	بنك جيد	1	0.064141	1
4	[0, 1.0165976300520323]	[0.7071067811865476, 0.7071067811865476]	علة سيئه	0	0.718843	1
5	[1.0099803626387285]	[1.0]	سيئ	1	1.009980	1
6	[-1.014209119433242, 0.9987544275742212, 1.011...	[0.2695555378363775, 0.3582465229239333, 0.332...	تطبيق رائع عملي شيء بسيطا ويعمل جيدا بالنسبه	1	0.419243	1
7	[-1.014209119433242, 0.9987544275742212]	[0.6012421369081836, 0.7990668888185024]	تطبيق رائع	1	0.188286	1
8	[-1.0185605179730195, 1.0085361926300995]	[0.7846083335651886, 0.619991744219274]	مربحه للغايه	1	-0.173887	0
9	[-1.006881361597804, 1.012423789587569]	[0.7071067811865476, 0.7071067811865476]	تعمل جيده	1	0.003919	1
10	[1.013351799510099, 1.0077162015668324, 1.0054...	[0.6352499504494206, 0.4993920701885275, 0.589...	يعمل جيد بالنسبه	1	1.739288	1

3. Evaluation

3.1.Expérimente

Les métriques choisies pour évaluer les performances du modèle était la précision, le rappel et le score F et l'accuracy. Ce modèle atteint une précision de 0,83. ce qui montre qu'il était bon pour discriminer les observations de sentiments négatifs par rapport aux positifs.
Le modèle a également atteint près de 78 % de rappel (ce qui signifie que 78 % de toutes les observations positives de l'ensemble de données ont été correctement classées comme positives),

	scores
accuracy	0.714286
precision	0.838710
recall	0.787879
f1	0.812500

3.2.Résultats

Si nous comparons ces résultats avec ceux obtenus par l'algorithme supervisé svm, la précision du modèle non supervisé est en fait supérieure au modèle supervisé, et la précision est inférieure au modèle supervisé, bien qu'il soit difficile de comparer.

	scores unsupervised method
accuracy	0.714286
precision	0.838710
recall	0.787879
f1	0.812500

	scores supervised method
accuracy	0.753425
precision	0.542857
recall	0.904762
f1	0.678571

Pour résumer, l'approche non supervisée a obtenu des résultats satisfaites (à mon avis), car sans l'utilisation de modèles pré-entraînés, et en fait aucune information préalable sur ce qui est positif ou négatif dans un texte donné, elle a obtenu des métriques assez élevées, nettement élevées que prévu à Aléatoire.

4. Conclusion

Dans ce travail, nous avons utilisé l'apprentissage automatique non supervisé pour détecter et regrouper les avis des utilisateurs à-propos la banque CIH au Maroc.
Pour détecter et regrouper les sentiments des avis. À cette fin, nous avons recueilli environ 251 commentaires. De plus, nous avons utilisé une implémentation de gensim de l'algorithme word2vec avec l'architecture CBOW, le clustering, les fonctionnalités et stratégies de prétraitement pour trouver les meilleurs modèles pour prédire l'étiquette de sentiment. Les résultats ont montré notre model atteint des résultats satisfaisants

5. Références

[1]. Kaggle, "Gensim Tutorial Word2Vec."
[2]. Rafał Wójcik," Unsupervised Sentiment Analysis ",
[3]. Word2vec embeddings.
[4]. Dhruvil Karani, "Introduction to Word Embedding and Word2Vec".