Département d'Informatique et de Recherche Opérationnelle Université de Montréal

IFT3335: AI – Hiver 2025 EX3

FAIT PAR:
Asmaa Zohra Skou, 20217232
Nada Alem, 20193740
Peter El-Hadad, 20206705

1 Partie 1

1.1 Analyse des résultats

Question 1 :

Quelle méthode (BoW ou TF-IDF) donne les meilleurs résultats ? D'après les données du tableau, c'est la méthode Bag of Words qui offre les meilleures performances globales. Pour chaque valeur de `max_features`, la précision (accuracy) obtenue avec BoW est systématiquement supérieure à celle de TF-IDF.

Pourquoi est-ce le cas dans ce problème de classification ? Le modèle BoW fonctionne mieux que TF-IDF dans la classification du SMS, car il conserve l'importance brute des mots, ce qui est utile dans des messages courts et répétitifs comme les spams. Chaque modèle a un fonctionnement différent et répond à des objectifs distincts. La structure simple et directe de BoW est plus adaptée pour ce problème pour ces raisons :

1. Nature du dataset: les messages du jeu de données SMS Spam Collectio sont généralement très courts. Dans ce contexte chaque mot joue un rôle important. Puisque le modèle BoW se contete de compter la fréquence d'apparition des mots, capte directement cette information sans la modifier. En général, cela rend BoW particulièrement efficace sur des textes ou la simple présence ou répétition d'apparition des mots suffit à indiquer sa pertinence.
2. La fréquence des mots ont un sens dans le spam.
3. TF-IDF est plus utile dans d'autres contextes : pour des documents plus longs (articles, commentaires, emails) ou il faut différencier les mots importants des mots banals.

Source : Bag of Word and Frequency Count in Text using Sklearn.

Question 2 :

Que se passe-t-il pour les métriques lorsque l'on diminue `max features` ? Afin d'observer l'effet de la taille du vocabulaire sur la performance des modèles, nous avons fait varier la valeur de `max features` de 500 à 5000. Comme le montre le tableau ci-dessous, la précision (accuracy) du modèle BoW reste relativement stable, même avec un vocabulaire réduit. En revanche, la méthode TF-IDF montre une légère baisse de performance lorsque `max features` augmente, probablement à cause de l'ajout de mots moins discriminants.

Ces résultats indiquent que pour des messages courts comme les SMS, un vocabulaire réduit (par exemple 1000 ou 2000 mots) est suffisant pour maintenir de bonnes performances, tout en réduisant le coût de calcul.

Du côté de TF-IDF, on note une petite baisse des performances quand le nombre de mots augmente. Ce test nous a donc permis de mieux comprendre l'effet de ce paramètre et que l'ajuster en fonction des besoins et des ressources est une bonne idée.

Table 1: Performances selon la taille du vocabulaire (max_features)

max_features	Accuracy BoW	Accuracy TF-IDF
500	0.9767	0.9709
1000	0.9776	0.9718
2000	0.9795	0.9689
3000	0.9799	0.9664
4000	0.9795	0.9630
5000	0.9794	0.9614

2 Partie 2

2.1 Analyse des résultats

Table 2: Résultats globaux des trois techniques de vectorisation

Technique	Modèle	Accuracy	F1-score
BoW	Logistic Regression	0.97936	0.91826
	Random Forest	0.97038	0.88357
	MLP	0.98331	0.93508
TF-IDF	Logistic Regression	0.96141	0.83320
	Random Forest	0.97236	0.88328
	MLP	0.98098	0.92786
SBERT	Logistic Regression	0.98169	0.92818
	Random Forest	0.96985	0.88242
	MLP Classifier	**0.98618**	**0.94517**

Question 1 :

Comment les embeddings se comparent-ils à BoW et TF-IDF en termes de performance ?
Les embeddings SBERT donnent de meilleurs résultats que BoW et TF-IDF, surtout en termes de F1-score. Cela s'explique par le fait qu'un embedding pré-entraîné comme SBERT capture le sens global des phrases, et pas seulement la fréquence des mots. Contrairement à BoW ou TF-IDF, qui sont des méthodes purement statistiques, SBERT utilise le contexte et la structure linguistique, ce qui améliore la capacité du modèle à faire la distinction entre spam et ham même lorsque les mots sont similaires mais utilisés différemment.

Question 2 :

Quels sont les avantages et inconvénients d'utiliser des embeddings pré-entrainés ?

Avantages :

1. *Transformers employ attention mechanisms to excel in understanding relationships between distant words in a sentence.* Source : Transformers in NLP:BERT and Sentence Transformers. Cela nous dit que contrairement à BoW et TF-IDF qui ignorent l'ordre et le contexte des mots, les modèles comme Sentence transformers capturent efficacement les relations à longue distance grâce au mécanisme d'attention. Cela leur permet de mieux comprendre le sens global des phrases, là où les approches classiques se limitent à la simple fréquence des termes !

2. L'entraînement de word embeddings à partir de zéro est une tâche très exigeante en temps et en ressources. Une solution efficace consiste à utiliser des modèles pré-entraînés, déjà optimisés sur de grands datasets. Cette approche s'appuie sur le principe de *transfer learning*, qui permet de réutiliser un modèle existant pour l'adapter à une tâche spécifique. Source : Pre-Trained Word Embedding in NLP

Inconvénients :

1. Coût computationnel élevé : les modèles comme SBERT sont plus lourds à charger et à exécuter, ce qui les rend moins adaptés aux environnements avec peu de ressources.
2. Moins interprétables : contrairement à BoW ou TF-IDF, les vecteurs d'embeddings sont difficiles à analyser directement, car chaque dimension n'a pas de signification claire.
3. Dépendance au domaine du pré-entraînement : si les embeddings ont été entraînés sur un corpus différent (ex. Wikipédia), ils peuvent être moins performants sur un langage très spécifique (SMS, jargon professionnel, etc.).

3 Partie 3

3.1 Analyse des résultats

Table 3: Poids attribués par le méta-modèle à chaque modèle de base

Modèle de base	**Poids**
Logistic Regression	2.1443
Random Forest	6.8269
SVC	4.4417

Question 1 :

Observations : Les résultats obtenus avec le Super Learner montrent une performance globale très solide, avec une accuracy de 0.9835 et un F1-score de 0.9359, ce qui est comparable ou légèrement supérieur aux meilleurs résultats obtenus individuellement par les modèles BoW, TF-IDF et SBERT. Cela confirme que le Super Learner réussit à combiner efficacement les forces des modèles de base, en particulier en améliorant légèrement la stabilité des performances.

Question 2: Le méta-modèle a attribué des poids relativement élevés à Random Forest (6.83) et SVC (4.44), indiquant que ces deux modèles contribuent fortement à la décision finale. Le poids plus faible de LogisticRegression (2.14) montre qu'il est considéré comme utile mais moins déterminant.