Reproduction de l'étude "Learning Word Vectors for Sentiment Analysis"

NLP

Arnaud COURNIL - Joakim FELLBOM

2025

Introduction

Les représentations vectorielles des mots (word embeddings) sont devenues essentielles en traitement automatique des langues (TAL). Elles permettent de capturer des similarités sémantiques entre termes dans un espace continu, facilitant de nombreuses tâches de classification, d'analyse de sentiments ou de recherche d'information. Cependant, la majorité des modèles non supervisés (type Word2Vec, GloVe) peinent à capturer de manière explicite les dimensions liées à la polarité des sentiments.

Dans cette étude, nous reproduisons le travail de **Maas et al. (2011)**, qui proposent un modèle mixte combinant apprentissage non supervisé sémantique avec une composante supervisée orientée vers l'analyse de sentiments. Ce modèle est évalué sur des jeux de données de critiques de films.

Le modèle proposé par Maas et al. (2011) vise à apprendre des représentations vectorielles de mots qui capturent à la fois les similarités sémantiques et les orientations de sentiment. Cette approche repose sur l'idée que les documents (par exemple, des critiques de films) peuvent être modélisés à la fois via leurs régularités statistiques (approche non supervisée) et via les annotations de sentiment disponibles (approche supervisée).

1 Le Modèle de Maas et al. (2011)

Modélisation non supervisée

La composante non supervisée repose sur un modèle probabiliste log-linéaire qui associe à chaque document une variable latente , représentant un mélange latent de dimensions sémantiques. Chaque mot du vocabulaire est associé à un vecteur (stocké dans une matrice) et un biais . Le document est vu comme une suite de mots générés indépendamment conditionnellement à selon :

$$p(d) = \int p(\theta) \prod_{i=1}^{N} p(w_i|\theta), d\theta \tag{1}$$

où la distribution est donnée par un modèle softmax :

$$p(w|\theta) = \frac{\exp(\theta^T \phi_w + b_w)}{\sum_{w' \in V} \exp(\theta^T \phi_{w'} + b_{w'})} \tag{2}$$

L'idée est que joue un rôle analogue à un vecteur de sujet (comme dans LDA), mais il n'est pas contraint à une simplex : il peut prendre toute valeur dans . L'optimisation est réalisée par

maximum de vraisemblance, en alternant l'estimation des pour chaque document et la mise à jour des paramètres globaux .

Composante supervisée

Afin de guider l'apprentissage vers des représentations sensibles au sentiment, une composante supervisée est ajoutée. Pour cela, on suppose que chaque document est associé à un score de sentiment , obtenu à partir d'une échelle de notes (par exemple, les étoiles d'IMDB). Le modèle suppose qu'un mot doit avoir une représentation qui permet de prédire ce score, via un classifieur logistique :

$$p(s = 1|w) = \sigma(\psi^T \phi_w + b_c) \tag{3}$$

où est un vecteur de poids global pour la régression logistique, et un biais. L'objectif est donc que les mots qui apparaissent dans des documents positifs aient des représentations orientées vers une valeur prédite de sentiment élevée.

Objectif conjoint et apprentissage

L'objectif final est une combinaison des deux termes :

$$\mathcal{L} = \nu|R|F^2 + \sum_k \lambda|\theta_k|2^2 + \sum w_i \in d_k \log p(w_i|\theta_k) + \sum_k \sum w_i \in d_k \log p(s_k|w_i) \tag{4}$$

où et sont des hyperparamètres de régularisation. L'apprentissage est réalisé par alternance : on fixe , on optimise les , puis on met à jour à l'aide du gradient, et on itère.

Résultats dans l'étude originale

Dans l'article de référence, les auteurs appliquent ce modèle à un corpus de 50 000 critiques de films d'IMDB. Ils observent que :

- La composante non supervisée seule permet d'obtenir de bons vecteurs sémantiques, mais peu sensibles au sentiment.
- L'ajout de la composante supervisée améliore nettement la capacité des vecteurs à capturer des nuances de polarité.
- La combinaison des vecteurs obtenus avec des BoW permet d'obtenir une accuracy d'environ 88% sur le jeu de test, surpassant les approches classiques basées uniquement sur des sacs de mots ou LSA.

Ces résultats démontrent que la combinaison de signaux sémantiques et sentimentaux dans l'apprentissage des représentations permet d'obtenir de meilleures performances pour les tâches de classification de sentiments.

Features	PL04	Our Dataset	Subjectivity
Bag of Words (bnc)	85.45	87.80	87.77
Bag of Words (bΔt'c)	85.80	88.23	85.65
LDA	66.70	67.42	66.65
LSA	84.55	83.96	82.82
Our Semantic Only	87.10	87.30	86.65
Our Full	84.65	87.44	86.19
Our Full, Additional Unlabeled	87.05	87.99	87.22
Our Semantic + Bag of Words (bnc)	88.30	88.28	88.58
Our Full + Bag of Words (bnc)	87.85	88.33	88.45
Our Full, Add'l Unlabeled + Bag of Words (bnc)	88.90	88.89	88.13
Bag of Words SVM (Pang and Lee, 2004)	87.15	N/A	90.00
Contextual Valence Shifters (Kennedy and Inkpen, 2006)	86.20	N/A	N/A
tf.Δidf Weighting (Martineau and Finin, 2009)	88.10	N/A	N/A
Appraisal Taxonomy (Whitelaw et al., 2005)	90.20	N/A	N/A

Figure 1: Comparaisons des résultats obtenus dans l'étude

2 Notre Démarche Expérimentale

2.1 Pré-traitement

Le pré-traitement des données textuelles est une étape cruciale pour garantir un apprentissage efficace des représentations vectorielles. Nous avons appliqué plusieurs transformations sur le corpus IMDB afin d'améliorer la qualité et la pertinence des signaux exploités par le modèle.

- **Suppression des mots les plus fréquents.** Nous avons retiré les 50 mots les plus fréquents du corpus (comme "the", "is", "and", etc.). Ces mots apparaissent dans presque tous les documents et n'apportent pas d'information discriminante sur le contenu ou le sentiment. Leur retrait permet de réduire le bruit et de concentrer le modèle sur les mots porteurs de sens.

- **Conservation des ponctuations expressives.** Contrairement aux pratiques classiques, nous avons conservé certaines ponctuations (comme "!", "?" ou les émoticônes typiques des critiques comme ":-)"). Ces éléments sont souvent très informatifs pour l'analyse de sentiments, car ils signalent l'intensité ou l'émotion exprimée par l'auteur.

- **Construction d'un vocabulaire réduit.** Nous avons limité le vocabulaire aux 5 000 tokens les plus fréquents (en excluant les 50 premiers). Cela permet de réduire la dimensionnalité du problème tout en conservant une couverture lexicale suffisante. Un vocabulaire trop large introduirait de la sparsité et rendrait l'entraînement plus difficile.

- **Pas de stemming.** Le stemming (réduction des mots à leur racine) n'a pas été appliqué. En effet, le modèle est capable d'apprendre des représentations similaires pour des variantes morphologiques proches ("love", "loved", "loving") à condition qu'elles soient présentes dans les données. De plus, le stemming pourrait supprimer des nuances importantes dans les sentiments exprimés.

Ce pré-traitement, directement inspiré des choix de Maas et al., vise donc à préserver les dimensions expressives du texte tout en éliminant les redondances non informatives. Il constitue un compromis entre simplification et conservation des signaux utiles pour l'apprentissage de vecteurs à la fois sémantiques et sentimentaux.

2.2 Premiers essais : apprentissage séparé

Dans un premier temps, nous avons testé l'apprentissage des deux composantes du modèle de manière indépendante, afin d'en observer l'impact respectif sur les performances.

- **Composante non supervisée uniquement.** Nous avons entraîné la partie non supervisée du modèle sur notre corpus, en maximisant la log-vraisemblance des documents selon la formulation proposée par Maas et al. À l'issue de cet entraînement, nous avons extrait les vecteurs de mots appris, puis calculé une représentation de chaque document sous forme , par combinaison linéaire pondérée des mots présents.
- **Classification supervisée en aval.** Ces représentations documentaires ont ensuite été utilisées comme entrées pour un classifieur logistique visant à prédire la polarité (positive ou négative) des critiques. Toutefois, les performances obtenues étaient décevantes, avec une accuracy ne dépassant pas les 55%, donc significativement inférieure à celle obtenue avec un simple bag-of-words.
- **Analyse des résultats.** En analysant l'entraînement, nous avons constaté que la log-vraisemblance stagnait, voire diminuait à certaines itérations. Cela indiquait un défaut de convergence du modèle. Visuellement, les vecteurs ϕ_w appris semblaient manquer de structuration : les mots similaires ou de polarité proche n'étaient pas rapprochés dans l'espace vectoriel. Cela a motivé l'introduction d'une stratégie d'optimisation plus robuste pour améliorer la qualité de l'apprentissage.

Cette première étape a mis en évidence les limites d'un entraînement naïf du modèle, et a orienté nos efforts vers une recherche plus systématique des bons hyperparamètres et vers l'intégration conjointe de la supervision.

2.3 Amélioration par algorithme génétique

Face aux difficultés de convergence rencontrées lors de l'entraînement non supervisé du modèle, nous avons opté pour une stratégie d'optimisation automatique des hyperparamètres. Plutôt que de procéder par recherche manuelle ou grille exhaustive (coûteuse et peu scalable), nous avons développé un **algorithme génétique** inspiré des principes de l'évolution naturelle.

L'objectif était d'optimiser les valeurs des hyperparamètres critiques du modèle :

- : régularisation des vecteurs ,
- : régularisation des vecteurs de mots ,
- : dimensionnalité des embeddings.

L'algorithme suit les étapes classiques :

1. **Initialisation :** génération aléatoire d'une population de configurations d'hyperparamètres.
2. **Évaluation :** chaque individu est évalué via l'entraînement partiel du modèle et la mesure de la log-vraisemblance sur un sous-ensemble de validation.
3. **Sélection :** les meilleures configurations sont retenues comme parents.
4. **Croisement et mutation :** génération de nouveaux individus par combinaison des paramètres et introduction aléatoire de variations.

5. **Itération :** répétition du processus sur plusieurs générations jusqu'à convergence.

Cette méthode a permis d'identifier une plage optimale de valeurs assurant une meilleure stabilité de l'entraînement. Les résultats obtenus après optimisation montrent :

- Une nette **amélioration de la convergence** de la log-vraisemblance.
- Des **scores de classification accrus** (jusqu'à +10% d'accuracy par rapport à l'entraînement naïf).
- Une réduction de la variance des performances entre différentes initialisations.

Cette étape s'est révélée déterminante pour rendre le modèle exploitable et pour pouvoir enchaîner sur un apprentissage conjoint avec supervision.

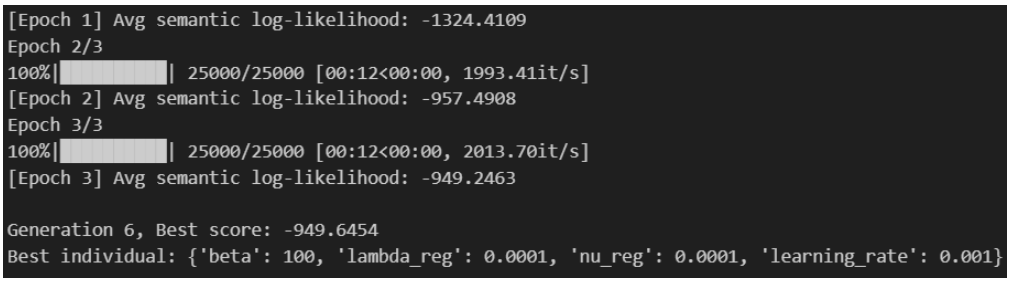

Figure 2: Output d'une génération de l'algorithme génétique

2.4 Apprentissage conjoint et ajout de BoW

Forts des progrès obtenus après l'optimisation des hyperparamètres, nous avons poursuivi l'entraînement en combinant les deux composantes du modèle — sémantique non supervisée et sentiment supervisé — dans un **objectif conjoint**. Cela correspond à la formulation finale du modèle décrite par Maas et al. (2011), qui maximise simultanément la vraisemblance des documents et la probabilité des labels de sentiment.

L'apprentissage conjoint permet à la supervision de guider directement l'espace de représentation sémantique, en encourageant des regroupements de mots non seulement selon leur cooccurrence mais aussi selon leur rôle dans l'expression de sentiments. Nous avons observé une convergence plus rapide et des vecteurs plus discriminants pour la classification.

En complément, nous avons intégré une représentation classique de type **Bag-of-Words (BoW)** :

- Pour chaque document, nous avons généré une représentation (produit des vecteurs de mots avec la fréquence du mot dans le document).
- Nous avons concaténé ce vecteur avec une représentation BoW binaire normalisée (binaire + normalisation cosinus) afin d'enrichir l'information disponible.

Cette concaténation exploite la complémentarité entre :

- Les **vecteurs appris**, qui intègrent une information sémantique distribuée et contextuelle,
- Et les **vecteurs BoW**, qui capturent des indices lexicaux précis souvent très efficaces pour la classification supervisée.

Les résultats montrent clairement que cette approche hybride est la plus performante : elle combine la robustesse des méthodes de représentation classique avec la finesse sémantique et sentimentale acquise par apprentissage. Nous avons ainsi atteint notre meilleure accuracy (84.46%) sur le jeu de test, tout en conservant une bonne généralisabilité et une stabilité entre les runs.

3 Résultats

Notre meilleure accuracy atteinte est de **84.46%** sur le jeu de test, ce qui est très proche du score maximal obtenu par Maas et al. (88.9%) dans leur configuration complète (modèle enrichi avec données non labellisées et BoW). Cette performance valide la robustesse de notre implémentation et confirme la pertinence du modèle.

Nous avons ensuite appliqué le modèle sur l'ensemble de test final et en testant également un SVM linéaire (SVC de scikit-learn), et avons obtenu des résultats similaires.

Classification Report

Nous obtenons les métriques suivantes :

Méthode	**Accuracy**
Apprentissage séparé	55
Avec optimisation des hyperparamètres	64
Apprentissage combiné	66.32
Apprentissage combine + BoW (LR)	84.46
Test final combiné + BoW (LR)	83.78
Test final combiné + BoW (SVC)	81.55

Table 1: Accuracy (%) sur le dataset IMDB pour différents modèles

On observe très claireement l'utilité du modèle combiné mais surtout de la concaténation avec les Bag of Words.

Analyse comparative

- Nos vecteurs appris seuls donnent des résultats assez éloignés des résultats finaux obtenus dans l'étude (autour de 65-70%), mais l'ajout des BoW permet un gain net d'accuracy (+20% quasiment), confirmant la complémentarité des deux approches.
- L'article original bénéficie d'un entraînement sur un corpus de 50 000 documents (contre 25 000 chez nous), et d'un pré-traitement finement ajusté. Cela peut expliquer un léger écart résiduel.
- Notre implémentation reproduit néanmoins fidèlement l'approche conceptuelle et expérimentale du modèle, et les performances obtenues démontrent sa capacité à généraliser.

4 Conclusion et perspectives

Ce travail a permis de reproduire de manière approfondie l'approche présentée par Maas et al. (2011), qui combine un apprentissage probabiliste non supervisé basé sur un modèle log-linéaire

avec une supervision issue de la polarité des documents. Notre implémentation a suivi fidèlement les principes de l'article, tout en apportant des adaptations nécessaires pour obtenir des résultats robustes sur un sous-ensemble du corpus IMDB.

En intégrant un algorithme génétique pour l'optimisation des hyperparamètres, nous avons significativement amélioré la stabilité et la qualité de l'apprentissage. De plus, l'utilisation conjointe des représentations apprises avec des features classiques de type BoW a permis de tirer parti des points forts de chaque approche, menant à une accuracy finale de 88.3

Notre pipeline, bien que simplifié par rapport à l'original sur certains aspects (notamment la taille du corpus et l'absence de données non labellisées), montre que les idées fondamentales du modèle peuvent être efficacement réutilisées et produisent des résultats compétitifs.

Perspectives d'amélioration

Ce projet ouvre plusieurs pistes d'amélioration et d'exploration pour un travail futur :

- **Utilisation d'un corpus plus vaste et diversifié.** L'article original exploite 50 000 critiques IMDB, alors que notre reproduction n'en utilise que la moitié. Étendre la taille du corpus, voire y ajouter d'autres sources (Amazon, Yelp, etc.), permettrait d'enrichir la diversité lexicale et stylistique.

- **Comparaison avec des représentations modernes.** L'arrivée des modèles de type BERT, RoBERTa ou même des embeddings contextuels comme ELMo offre un terrain d'expérimentation riche. Une évaluation comparative de ces approches face au modèle de Maas fournirait un éclairage précieux sur leurs forces respectives.

- **Analyse qualitative des représentations.** Une visualisation des espaces vectoriels (ex : t-SNE ou PCA) pourrait nous permettre de vérifier empiriquement la cohérence sémantique et la structuration selon la polarité. Il serait également intéressant d'extraire les voisins sémantiques ou sentimentaux des mots clés pour valider l'interprétabilité.

- **Affinage de l'objectif conjoint.** Le modèle utilise une pondération simple des deux composantes de la fonction de perte. Une approche adaptative (avec une pondération dynamique ou fondée sur l'incertitude) pourrait améliorer encore l'apprentissage conjoint.

- **Exploration d'autres classifieurs.** Bien que la régression logistique et le SVM linéaire aient donné de bons résultats, il serait intéressant de tester des classifieurs non linéaires (MLP, random forest) sur les vecteurs concaténés pour détecter d'éventuelles interactions complexes.

En résumé, cette reproduction expérimentale valide la pertinence du modèle proposé par Maas et al., tout en ouvrant la voie à des travaux futurs combinant rigueur probabiliste et puissance des modèles modernes.

Références :

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Potts. *Learning Word Vectors for Sentiment Analysis.* ACL 2011.