Évaluation de la positivité d'un tweet à l'aide des LSTM

Cédric Grelier & Lorent Caravaku

INF8225 - Projet de session
Polytechnique Montréal

12 Avril 2018

Introduction

- Travaux connexes
- Approche théorique & expériences
- Résultats & analyse
- Conclusion

Travaux connexes

- Nombreux travaux depuis les années 2000
- Dans les années 2000 : SVM, classification naïve bayésienne (Bo Pang 2002) et méthodes avec lexiques de mots (Xiaowen Ding 2008).
- 2015 : Premier travail avec LSTM (Duyu Tang 2015)
- Dans les années 2010 : Concours sur l'évaluation de sentiment (SemEval ; 88,2%)
- Sentiment140 (Alec Go 2009) : 1,5 Millions de Tweets

Corpus de données

Données: tweets sur sujets quelconques datant de 2009
Labels: *négatif* et *positif* + *neutre* dans le corpus de test

Entraînement	Test
1 599 961	359

Table 1: Nombre d'éléments par corpus de données

Créé automatiquement à l'aide des smiley (Go, Bhayani et Huang, 2009)

Nettoyage des données

- hashtags (#exemple)
- urls (http://www.exemple.com)
- liens des noms d'utilisateurs (@exemple)

Fichier passe de 233 Ko $\rightarrow$ 123 Ko

Prolongement de mots

Associer une représentation vectorielle numérique à un mot

- Word2Vec
- GloVe
- ...

Pré-entraînement

Fichier glove.twitter.27B.50d
(`https://nlp.stanford.edu/projects/glove/`)
Un mot $\rightarrow$ vecteur de dimension 50
Entraîné sur 2B de tweets contenants 27B mots distincts

Modèles

Modèle	Couches
1	Embedding (Pre-embedding) Conv1D-128 Conv1D-64 Conv1D-32 MaxPooling1D Dropout Bidirectional LSTM-128 Bidirectional LSTM-128 Dropout Dense
2	Embedding (Pre-embedding) Conv1D-256 MaxPooling1D Dropout LSTM-128 Dropout Dense
3	Embedding (NON Pre-embedding) Conv1D-64 MaxPooling1D Dropout Bidirectional LSTM-128 Dropout Dense
4	Embedding (Pre-embedding) Conv1D-64 MaxPooling1D Dropout Bidirectional LSTM-128 Dropout Dense
5	Embedding (Pre-embedding) Dropout Bidirectional LSTM-64 Bidirectional LSTM-64 Dropout Dense
6	Embedding (Pre-embedding) Dropout Conv1D-256 MaxPooling1D Dropout LSTM-128 Dropout Dense
7	Embedding (Pre-embedding) Dropout Conv1D-64 MaxPooling1D LSTM-70 Dense

Figure 1: Nos 7 modèles

Expériences

- Données : Sentiment140 : 1,5 Millions de Tweets
- Early-stopping : sur la précision sur les données de validation et avec une patience de 3
- Données de Validation : *validation_split* : 0,2

Résultats

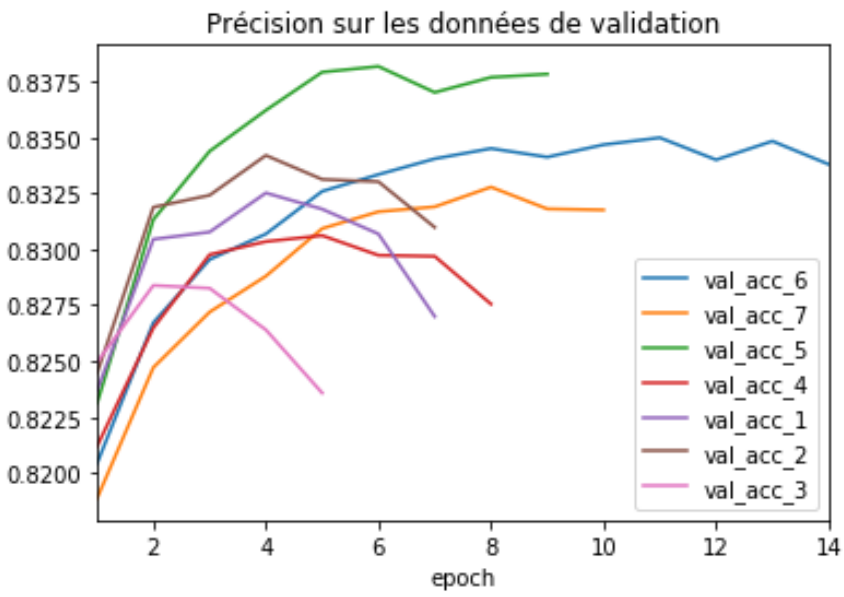

Figure 2: Précision

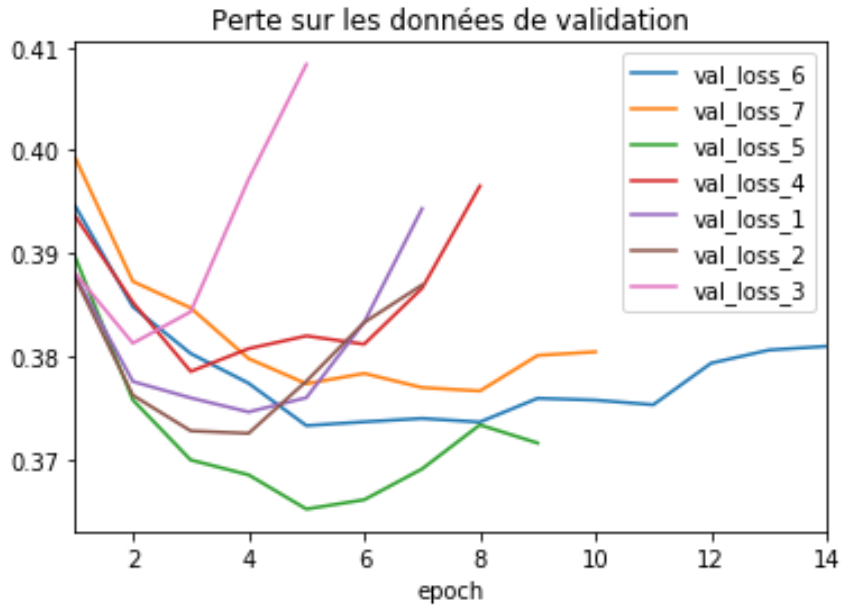

Figure 3: Perte

Résultats

Modèle	Précision	Perte	Temps par epoch (s)
1	0.8050	0.4469	3 000
2	0.7966	0.4626	**1 390**
3	0.8161	0.4283	1 450
4	0.8105	0.4378	1 450
5	0.7994	0.4270	10 200
6	0.8161	0.3993	1 450
7	**0.8272**	**0.3954**	1 550

Table 2: Précision et perte obtenues sur l'ensemble de test pour les différents modèles

Analyse

- Impact de l'utilisation d'un pré-entraînement
- Utilisation d'un ensemble de données de test construit différement du corpus d'entraînement
- Importance de la configuration du modèle sur le temps de calcul par epoch

Conclusion

Résultats proches de Go, Bhayani et Huang en 2009
Pas d'amélioration notable grâce au LSTM
À relativiser étant donné le peu de ressources mises en jeu

Améliorations

- Ajout d'une classe neutre
- Gestion des doublons du langage courant ("byyyyyye" et "bye")
- Exploration de paramètres nécessitants plus de puissance de calcul