Analyse de sentiment sur des critiques de films

Léa Abriel, Annah Augier, Anne Coustans

April 2, 2024

1 Présentation des données et de la problématique

L'objectif du défi que nous avons choisi sur la plateforme Challenge Data est de prédire le sentiment (positif ou négatif) de critiques de films en anglais. On a ainsi en données d'entrée 10 000 critiques de films, qu'il faut nettoyer et avec des labels qui sont à corriger. Ainsi, l'objectif est d'augmenter l'accuracy, soit une métrique de performance, à partir du modèle donné par le challenge (fastText) en corrigeant les labels, passant ainsi par du nettoyage des données textuelles, des corrections orthographiques, lemmatisation et tokenisation.

2 Prétraitement des données

La première étape a été de nettoyer les données textuelles, tel que supprimer des caractères spéciaux, des balises HTML, des espaces multiples afin de faciliter la tokensisation entre autre. Ensuite, nous avons selectionné un modèle pour corriger les fautes d'orthographe en utilisant jamspell, qui était beaucoup rapide et aussi efficace qu'utiliser textblob. Enfin, nous avons procédé à une lemmatisation, permettant de ramener les mots à leur forme radicale. Cette technique est plus précise car elle prend aussi en compte le contexte du mot contrairement au stemming, qui essaie de trouver la racine du mot sans tenir compte du contexte donc qui perd en efficacité. Nous avons également éliminé les stop words, qui sont les mots courants dans la langue et qui n'apporte aucun élément à notre analyse de sentiment.

A partir de ce prétraitement des données, et après un division de la base entre test et en entraînement, nous avons testé le modèle utilisé pour ce challenge. Il s'agit d'un modèle de classification de texte en utilisant fasttext avec comme paramètre 0.1 en taux d'apprentissage, 20 itérations et 2 n-grams de mots pour l'entraînement. L'accuracy augmente alors au plus de traitements sont réalisés sur les avis de films. Elle passe alors de 0.5995 avec les données par défaut, à 0.6105 pour les données nettoyées de caractères spéciaux, à 0.6275 pour les avis corrigés de fautes d'orthographe et enfin 0.636 pour les données lemmatisées.

3 Amélioration de l'analyse de sentiment

Le but, après avoir testé le modèle naïf, est d'améliorer nos performances en testant de nouveaux traitements. Pour cela, nous avons importé plusieurs modules à partir du Natural Language Toolkit. En l'occurence, nous avons utilisé wordnet pour comprendre la signification des mots et leurs synonymes, sentiwordnet qui ajoute des informations sur la polarité et la subjectivité des mots et enfin postag, fonction de NLTK pour l'étiquetage grammatical.

La fonction penntreebank_to_wordnet convertit les étiquettes grammaticales du format Penn Treebank (utilisé par NLTK) en format compréhensible par WordNet. Cette étape est cruciale car WordNet et SentiWordNet utilisent des étiquettes différentes de celles générées par pos_tag.

La tokenisation est cruciale dans un contexte de données textuelles et permet de diviser le texte (un avis sur un film dans notre cas) en token, représentant un bout de mot, un mot ou un groupe de mots selon le contexte. Ensuite, il faut réaliser un étiquetage grammatical pour chaque token (nom,

adjectif, ect.) en utilisant nltk.pos_tag

La prochaine étape a été de calculer le score de sentiment. Premièrement, nous avons initialisé à 0 pour chaque avis puis appliqué la fonction de conversion des étiquettes grammaticales en format compréhensible par WordNet pour chaque token, qui permet une compatiblité avec les fonctions de lemmatisation. A partir de cet étiquetage, on filtre les token pour ne garder seulement les adjectifs, les noms et les adverbes qui portent généralement les sentiments. On applique sur ces nouvelles données la lemmatisation pour ne garder que la forme radicale du mot.

Pour le lemme de chaque mot, la fonction cherche les synsets correspondants dans WordNet, en se limitant au tag WordNet pertinent. Si aucun synset n'est trouvé, le mot est ignoré. Un synset est une contraction de "set of synonyms" (ensemble de synonymes) et représente un concept lexical unique et composé d'une liste de mots ou expression considérés comme sémantiquement équivalent dans un certain contexte. Ces ensembles de synonymes sont ainsi utiliser dans notre analyse de sentiment pour identifier les connotations positives ou négatives associées à des mots ou expressions.

Nous calculons ensuite le sentiment en utilisant SentiWordNet. Pour le premier synset de chaque mot (le plus commun), on calcule en utilisant SentiWordNet fournissant des scores positifs ou négatifs pour chaque synset. Le score de sentiment global pour le mot est calculé en soustrayant le score négatif du score positif. On cumule ensuite les scores pour en avoir un global pour l'avis. Ainsi, un avis avec un score positif global élevé est considéré comme positif, tandis qu'un score négatif indique un avis négatif.

Cette méthode permet d'obtenir une évaluation quantitative du sentiment exprimé dans chaque avis, en se basant sur les contributions positives ou négatives des mots pertinents. Avec ce modèle, nous obtenons une accuracy bien plus élevée que seulement sur nos données nettoyées. En effet, l'accuracy est passée de 0.636 avec les données lemmatisées, à 0.794 avec cette méthode de calcul de score de sentiment, ce qui est une grande amélioration.