Victoria Beltrán Domínguez

Exercice 2. Sentiment Analysis

Introducción

En esta práctica se tiene como objetivo la creación de un modelo que clasifique la polaridad de un tweet. Para ello, se realizarán diferentes pruebas, expuestas en el apartado de experimentación realizada y se terminará con unas conclusiones.

Experimentación realizada

Se han probado diferentes modelos con el fin de clasificar la polaridad de los tweets. A continuación se muestra una tabla con los experimentos realizados y los resultados obtenidos, dónde el enfoque básico representa aquel introducido en la ayuda de la práctica, con Count Vectorizer para sacar las features (utilizando el tokenizador presentado en la ayuda) y support vector machine (LinearSVC(C=0.1)) para entrenar el clasificador.
En las diferentes pruebas realizadas, se han cambiado elementos de este modelo. Si no se especifican diferentes, las técnicas por defecto serán las mencionadas.
En caso contrario, pasamos a indicar el significado de algún cambio en el modelo:

1) Tokenizador propio: aquel realizado para la práctica previa.

El resto de pruebas se entiende que el lector entiende el cambio realizado.

Enfoque	macro precision	macro recall	macro f1-score	micro accuracy
Básico	0.41	0.40	0.39	0.55
Básico + Conteo polaridades de palabras	0.39	0.39	0.39	0.53
Básico + tokenizador propio	0.43	0.41	0.40	0.56
Básico + tokenizador propio + Count Vectorizer con bigramas	0.40	0.32	0.29	0.48
Básico + HashingVectorizer	0.31	0.30	0.27	0.47
Básico + tokenizador propio + TfidfVectorizer	0.28	0.35	0.30	0.55
Básico + tokenizador propio + Count Vectorizer + SVM	0.11	0.25	0.15	0.43
Básico + tokenizador propio + StandardScaler + LinearSVC	0.38	0.36	0.35	0.50
Red neuronal con una capa densa que utiliza embeddings preentrenados	0.28	0.36	0.31	0.55
Count Vectorizer+ Stochastic Gradient Classifier	0.37	0.37	0.37	0.47
Modelo por votación*	0.48	0.35	0.33	0.53

(*) Modelo por votación: sistema que hace uso de diferentes modelos con el fin de conseguir una clasificación por votación. Se han utilizado tres modelos:

a) Modelos básico + tokenizador propio.
b) Modelo de red neuronal con una capa densa que utiliza embeddings pre entrenados.
c) Modelo básico + tokenizador propio + StandardScaler + SGDClassifier.

En caso de empate se clasifica el tweet según el modelo básico, pues es el que mejores resultados ha dado.

Conclusiones

En esta práctica se han probado diferentes modelos de clasificación de tweets. Se ha probado a añadir conteos de polaridades de palabras en la matriz de features. Se ha probado a utilizar una red neuronal con una capa densa que utiliza embeddings pre entrenados. Se han probado diferentes métodos de vectorización y de entrenamiento. Incluso se ha probado un modelo de votación, en el que se combinaban diferentes modelos con el fin de mejorar el resultado. Y sin embargo, el modelo que mejor ha funcionado ha sido el básico con el tokenizador propio. No encuentro una explicación clara de este suceso pero una posible hipótesis podría ser que el básico consigue correctamente procesar la frase en un vector que es realmente representativo, y por el que es más fácil aprender un modelo lineal.