Análisis de sentimientos en Twitter sobre Elon Reeve Musk en español, Regularizando Y Usando el Análisis de componentes principales como método de minimizador de dimensiones, en el 29/6/2022

Matias Serena

Colegio Universitario IES Siglo 21

Informe realizado en la carrera de inteligencia artificial para la asignación de procesamiento de lenguaje natural

matiaserena@gmail.com

Abstract

En el Procesamiento del Lenguaje cada palabra que utilizamos tiene un peso y un sesgo en las personas que puede ser positivo como negativo que nos puede decir que piensan la gente sobre un tema. Así podríamos hacer un análisis de los sentimientos sobre que piensa la gente sobre un tópico. En este trabajo se mostrará, cual modelo de Machine Learning es el mejor para el procesamiento de sentimientos sobre un tema elegido. En este caso será sobre Elon Musk con el agregado del uso del Análisis de componentes principales también llamado PCA

Keywords: procesamiento de lenguaje natural, análisis de sentimientos, Twitter, regularizado, vectorizado, PCA, castellano,

Marco Teórico

1.1 Procesamiento del Lenguaje Natural

Citando la definición de Antonio Moreno investigador de la UAM sobre que es el procesamiento del lenguaje natural [1]

” El Procesamiento del Lenguaje Natural es el campo de conocimiento de la **Inteligencia Artificial** que se ocupa de la investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino” parafraseando es como nos comunicamos a través de nuestro idioma con las maquinas utilizando la inteligencia artificial para lograrlo.

1.2 Inteligencia Artificial

Como anteriormente fue mencionado que el PLN es el campo de conocimiento de la inteligencia artificial parafraseando un poco, pero. ¿qué es inteligencia artificial? Para Juan Antonio Pascual Estapé es:

[2]“LA CAPACIDAD DE QUE LAS MÁQUINAS PIENSEN Y RAZONEN POR SU CUENTA [...] No existe un definición aceptada por todos los expertos de LO QUE SIGNIFICA LA INTELIGENCIA ARTIFICIAL. Primero, porque es una ciencia nueva, cambiante y experimental. Y segundo, porque ni siquiera podemos definir con exactitud qué es la inteligencia humana...”

Como se puede ver en esta cita es como un software en cierta forma puede usar algoritmos para la resolución de problemas y que es muy difícil de definir ya que no existe con exactitud una definición para qué es la inteligencia humana una de las problematicas

1.3 Análisis de sentimientos

Según Roberto Cuadros Muñoz profesor de lengua española, en la universidad de Sevilla la definición de análisis de sentimientos es [3]” El análisis del sentimiento que brindan las herramientas computacionales de procesamiento del lenguaje natural e inteligencia artificial nos permite comprobar los sentimientos y emociones en los medios digitales. Actualmente, es posible destilar la polaridad e intensidad mediante la identificación de rasgos léxicos, iconográficos y estructurales.” Parafraseando a través de las palabras y las forma de hablar se puede realizar una distensión si de lo que hablamos es positivo o negativo. Se puede ver que para investigar un problema de análisis de sentimiento va a tener una gran dimensionalidad

1.4 Análisis de componentes principales o Principal Component Analysis (PCA)

Una de las complejidades que tiene el modelo es la multidimensionalidad que tiene el problema para simplificarlo acudimos a la estadística en el uso del PCA pero primero lo definirá Joaquín Amat Rodrigo

[4]” Principal Component Analysis (PCA) es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables (X1, X2, …, Xp), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes (z<p) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal.”

Parafraseando el PCA a reducido el problema de multidimensionalidad en simple vectores que generalizan en los featuring menos importantes y guiando me en el paper de [5](Vikas Raunak, 2017) ”Simple and Effective Dimensionality Reduction for Word Embeddings” suele mejora los modelos de ML con esta reducción.

Marco Metodológico

Usando el leguaje de programación Python, notebook de jupyter y las librarías de pandas, skelearn, keras, numpy, matplotlib, twepy y tensorflow.

Para realizar este proyecto podemos dividirlo en 3 partes Scraping, limpieza de texto y entrenamiento de modelo

2.1 Scraping

Para el escrapeo de datos hemos utilizado la librería de Python twepy que nos permite conectarnos a la API de Tweeter a través de keys propias que anterior mente tuvimos que haber solicitado en la pag https://developer.twitter.com/en.

Los datos que serán escrapeados son fecha, tweets y usuario que estén en español y que hablen del tópico en nuestro caso de Elon Musk

Estos datos serán acondicionados para no tener emojis, ni links, ni hashtags, ni arrobados ya que no nos serán útil para el análisis del mismo y así serán almacenados en la base de datos. En nuestro caso usamos un base de datos de Oracle donde nos permitirá y categorizando que sentimiento despierta sobre tema que pueden se positivo, negativo o regular

2.2 limpieza de texto

importamos nuestro csv ya clasificados a manos los sentimientos de los twits. Hemos decido dropear los todos los datos menos los textos de los tweets y la clasificación asignada.

Para la limpieza de tokens minusculizamos todas las palabras y usamos regular expresión para no almacenar links, ni hashtags, ni arrobados ya que generarían ruidos en nuestro data frame. Lematizamos el texto en busca del lema de palabra para que palabras parecidas que significan lo mismo estén agrupadas cuando se hagan el Count Vectorizer que sirver para generar una matriz de números que le da peso a las palabras donde cada columna se una palabra diferente y las filas sean que sentimiento despierta la frase.

Al hacer el tener una columna por cada palabra usada para hablar del señor Musk nos vamos a dar cuenta la multidimensionalidad del problema, es decir la gran cantidad de columnas que usaremos para ello es que utilizaremos PCA y así reducirlo en simple 100 columnas de forma arbitrarias.

2.3 entrenamiento de modelo

se dividirá el Data Frame en X_train e Y_train y X_test e Y_test para comprobar cual fue el mejor modelo. Entre los modelos que se probaran están: Redes Neuronales, KNeighbors Classifier, Árbol, Regresión logística, Super Vector Classifier, Random Forest Classifier, Ada Boost Classifier, Extra Trees Classifier, Gradient Boosting Classifier

Resultado

El resultado de la investigación de los modelos comparado por el accurate nos dio

Comparativa de accurate entre los modelos de Machine Learning

modelos de ML PCA 100	AC
DNN	0.86
KNN	0.707
ARBOL	0.579
RL	0.671
SVC	0.66
RFC	0.69
AB	0.66
ETC	0.7
GBC	0.692

Conclusión

Los modelos de Redes Neuronales combinado con análisis de componentes principales demuestran que son las más eficaces para resolver los análisis de sentimientos y demuestra que gran potencial tiene este campo del Procesamiento del Lenguaje Natural.

Referencias

Moreno, A. (2018). Procesamiento del lenguaje natural ¿que es?.. *IIC*, https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/#:~:text=El%20Procesamiento%20del%20Lenguaje%20Natural,el%20ingl%C3%A9s%20o%20el%20chino.

Pascial Estapé, J. A. (2017). *Inteligencia artificial: qué es, cómo funciona y para qué se utiliza en la actualidad.* computerhoy.

Roberto Cuadros, M. (2022). ¿Puede la inteligencia artificial analizar los sentimientos y emociones de un tuit?. The conversation

Amat J, R. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. cienciadedatos

Vikas Raunak, (2017). Simple and Effective Dimensionality Reduction for Word Embeddings. https://arxiv.org/pdf/1708.03629v3.pdf