Analisis de sentimiento de textos financieros

Pablo Franco

Universidad Tecnologica Ncional

Analisis de sentimiento de textos financieros

Introduccion

En el marco de la cursada de la materia Procesamiento del Lenguaje Natural de la Universidad Tecnologica Nacional, se realiza esta actidad para fijar conocimientos mediante la aplicacion practica de tecnicas de analisis de sentimiento de textos y tecnicas de recuperacion de informacion.

Objetivo

Clasificar un texto para identificar si el sentimiento que un lector interpreta es positivo, negativo o neutro dentro de un contexto financiero. Esta informacion tiene como finalidad ayudar a analizar el comportamiento de activos financieros. Se utilizo como muestra activos del mundo de las criptomonedas.

Tecnologias

Se utiliza para la realizacion del proyecto una implementacion del lenguaje de programacion Python 3 por ser esta tecnologia la mas desarrollada y extendida en los campos academicos de la ciencia de datos. La implementacion utilizada en esta practica se denomina CPython version 3.10. Esta implementacion se puede utilizar y descargar libremente de internet. Python 3 es compatible con gran variedad de arquitecturas y sistemas operativos, en este ensayo mas especificamente se utilizaron arquitecturas amd64/x86 sobre Windows 10 y Debian. Durante la ejecucion del programa se observa que la huella de memoria principal necesaria para procesar las noticias atraves del modelo predictivo se aproxima a los 8gb. Vale mencionar la utilizacion de 2 modulos de Python fundamentales para la realizacion de esta practica, HappyTransformer (modulo de python para aplicar modelos predictivos preentrenados) Newspapaper (modulo de python para extraer y normalizar el texto de una pagina web) y NewsApi (modulo de python que permite acceso a un concentrador de feeds actualizados de cadenas de noticias internacionales). Tambien se utilizo un modelo preentrenado de tipo BERT llamado finbert

elaborado a partir de la valoracion de sentimiento de textos del mundo financiero.

Metodo

Se escogen diferentes cadenas de noticias para utilizar como fuente para el analisis de sentimiento de textos. Ademas de las fuentes se escogen 20 nombres de criptomonedas de las cuales recuperar las noticias. Esta informacion se utiliza para obtener los noticias por medio del concentrador de noticias internacionales tomando todas las noticias relacionadas a partir del dia anterior hasta el momento actual. Estas noticias se normalizan por medio del modulo newspaper el cual utiliza modelos de pln para corregir la puntuacion y los simbolos o definiciones de html que no corresponde encontrarlos dentro del texto. Al texto normalizado se le aplica un modelo predictivo tipo BERT denominado finbert para clasificarlos en positivo, negativo, neutro. En base a estos resultados se hace un conteo de las probabilidades para obtener un estimado del sentimiento aproximado de cada criptomoneda.