28 de abril 2020

Dudas Memoria v1

Antes de todo, aquí está el repositorio de Github donde voy haciendo avances con mi tésis. He probado varias alternativas para estimar los retornos de los activos del mercado. Al final, la más exitosa ha sido usar una versión de BERT preentrenada para estimar el sentimiento de reseñas de productos.
A continuación viene una serie de dudas que me gustaría plantear:

1. **Contribución:** He enfocado la contribución del trabajo como dar respuesta a una pregunta: *"Can we predict more accurately the returns of a stock with GAN based model if we include newspaper text data into it?"* ¿Qué te parece?

2. **Tiempo verbales:** No sé muy bien como gestionar los tiempos verbales a la hora de escribir el texto. De momento lo he hecho como me ha ido saliendo, pero entiendo que debería haber un consenso. En la intro pongo las cosas en futuro, hablando de lo que "voy a hacer", pero no sé si tiene mucho sentido. Luego paso al pasado, "he hecho", etc… ¿Debería escribir toda la tésis en el mismo tiempo verbal?

3. **Personas**: También tengo dudas respecto a la persona que utilizar al describir lo que se va haciendo. No se si es mejor decir usar la primera del singular i.e "I trained the model", o la primera del plural: "We train the model", o escribir todo en pasiva (puede quedar bastante aburrido) "the model is trained". ¿Qué persona crees que es más apropiado utlilzar?

4. **Modelo inestable**: En la primera iteración del trabajo hago predicciones de los retornos de Bitcoin y Tesla sólo usando texto como input (ver sección 3.1 y 3.2). En resumen, paso texto de noticias por BERT, obtengo una predicción del sentimiento de cada frase, agrupo las predicciones por días haciendo la media y entreno una LSTM que estime los retornos directamente. El problema es que las predicciones del modelo LSTM son demasiado inestables, dependen demasiado del tamaño de la ventana escogida (número de días hacia atrás utilizado para estimar el valor siguiente). ¿Conoces alguna manera de estabilizar las predicciones de redes LSTM.

5. **Media del Sentiment analysis:** para combinar los datos de sentiment analysis de todas las frases que pertenecen al mismo día hago la media de sus vectores de sentimiento que me devuelve BERT. En la siguiente tabla se ejemplifica esto:

28 de abril 2020

Dataset con frases cuyo sentimiento ya ha predicho BERT

Frases	Fecha	Vector sentimiento
Frase 1	1	[1,0,0,0,4]
Frase 2	1	[2,0,0,0,1]
Frase 3	1	[3,0,0,0,4]
Frase 4	2	[1,1,1,1,1]
Frase 5	2	[3,3,3,3,3]

Dataset tras hacer la media de las frases por días

Fecha	Media de Vector sentimiento
1	[2,0,0,0,3]
2	[2,2,2,2,2]

No estoy seguro si estoy perdiendo mucha información al hacer esto, pero si quiero usar una LSTM necesito un único vector de sentimiento que represente cada día. ¿Se te ocurre alguna alternativa para agrupar los datos de sentiment analysis de otra forma? ¿Me olvido de la LSTM?