En la siguiente versión asgurate de añadir:

o Sección de trabajos relacionados con al menos 10 trabajos científicos explicados en el  
contexto de la contribución esperada y bibliografía incluyéndolos en formato IEEE.  
o Resumen dejando clara la contribución esperada (resumen en texto y en gráfico).  
o Título de todas las secciones y subsecciones que contendrá el TFG/TFM.

Das una sección de trabajos relacionados, pero en realidad son preliminares: lo que son las GANs o lo que es BERT. Consulta las siguientes secciones de la guía de TFMs que te he dejado en Trello:

* 2.4 Trabajos relacionados y bibliografía
* 2.5 Trabajos relacionados vs preliminares

Dudas Memoria v1

Antes de todo, [aquí](https://github.com/juanluisrto/stock-prediction-nlp) 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.

Estupendo, añade una licencia GPL y también debería figurar yo. Si quieres puedes poner algo como “este código fuente se ha hecho para el TFM titulado…. Autor: tú, Director: yo. Así se entiende que tu eres el autor principal pero yo también tengo propiedad intelectural.

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?

En principio bien, pero eso es algo que tienes que decidir tú y mojarte. El riesgo es que ya haya trabajos que hacen eso. Por eso la sección de trabajos relacionados y poner la contribución en el contexto de estos trabajos es importante.

Una vez más, consulta las siguientes secciones de la guía:

2.4 Trabajos relacionados y bibliografía

2.5 Trabajos relacionados vs preliminares

Yo creo que será fácil encontrar trabajos de GANs para pedecir bolsa, y será fácil encontrar trabajos que usen machine learning y datos de periódicos (por no decir que casi todos lo harán). Pero quizás no sea tan común hacer una evaluación de hasta que punto los datos de los periódicos importan. Fíjate que en este caso, quizás ni siquiera te interese usar GANs o BERT, sino partir de alguna propuesta rival que encuentres en la literatura.

Si por otro lado tu propuesta va más en la línea de “Consigo una precisión más alta porque uso noticias de periódicos”, tendrás que comparte con alguna propuesta rival en la literatura (porque usas mismos datos o parecidos y ellos dan resultados o porque tienes su código y lo has usado en tus datos)

Una vez más, este tipo de cosas se perfilan en los trabajos relacionados y te darán ideas sobre cómo venderlo. También te dará ideas de una buena estructura de la tesis y para añadir esas subsecciones que te pedía en la primera entrega.

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?

Puedes usar We (en el sentido de que somos autor y supervisor) pero también puedes usar el pasivo. En papers científcos yo uso lla pasiva. No hables en futuro nunca, salvo quizás en una sección de trabajos futuros, porque en el momento de presentar la tesis y alo habrás hecho. Por lo mismo, tampoco en pasado.

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?

Pasiva. Y si es de otras personas, revisa las secciones

2.4 Trabajos relacionados y bibliografía

2.5 Trabajos relacionados vs preliminares

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.

Por ahora me preocupa la contribución y si tiene sentido hablar de LSTM o BERT. ¿Con qué otras propuestas vas a comparar? ¿mides una precisión mayor en global o sólo la mejora que aporta usar noticias sobre un baseline?.

No caigas en el error de probar los tres millones de arquitecturas de Deep learning sin saber hacia dónde vas.

**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:

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?

Sí claro, cientos de formas. Entre otras pod´rias pasarle todas las frases concatenadas a ELMO o BERT y debe´ria devolverte un único vector, porque vale para palabras/frases/párrafos. También puedes pasar a la LSTM una secuencia de palabras (codificadas en BERT) y generará un estado para el día, o pasarle (vector de ) frases. O hacer medias de estados ocultos de la LSTM en lugar de coger ´solo la salida. O usar Bidirectional LSTM con todas las opciones anteriores. Como te digo es infinito. Pero en este punto centrate en definir cual va a ser la contribución y cómo la vas a evaluar porque a lo mejor no es prioritario usar Bert o LSTM como te ponía en notas anteriores.

Lo ideal es compararte con algo que esté ya hecho en la literatura/paper. Si no es posible, tendrías que dar mucha variedad de métodos y compararlos, partiendo de casos lo más simple posibles (si prentendes aportar mejor predicción de bolsa, entiendo que el análisis de series temporales clásico con ARIMA y vanilla Recurrent Neural Networks)