Analisis de sentimientos sobre comentario en twitter con identificación de topicos

Esteban Jaramillo
ejaram11@eafit.edu.co
School of Applied Sciences and Engineering,
Universidad EAFIT, Medellín, Colombia

1 Descripcion del proceso

Para la elaboracion de este trabajo seguimos el siguiente proceso:

Se importaron los datos y se realizo un preprocesamiento de los mismos que contaba con lo siguiente:

1. se limpiaron las palabras de caracteres especiales y numeros.
2. se removieron las stop words utilizando la libreria de nltk.
3. se realizo la lematizacion de los datos.
4. se creo una matriz de term frecuency para futuro analisis de los datos .
5. se creo una matriz tf idf para el entrenamiento de los modelos.

2 EDA

Los datos estan compuestos por 5842 twitts en ingles los cuales fueron seleccionados entre el 2007 y 2008 y agrupan twitts que miden el sentimiento y la percepcion de la economia en ese periodo del tiempo Los twitts tienen una longitud de entre 7 y 298 caracteres contando espacios donde el twitt promedio constaba de aproximadamente 109 caracteres. Los tags que se les dieron a estos estan divididos en 3: neutral, positivo y negativo.

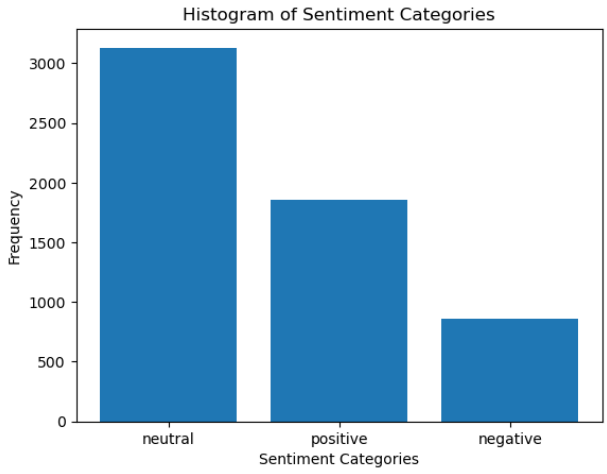

Figure 1: Distribucion Sentimientos

como podemos observar en la grafica, el sentimiento dominante es neutral, seguido por sentimientos positivos y finalmente negativos.

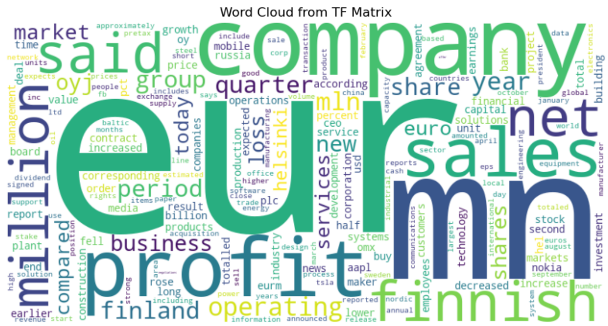

Figure 2: Palabras

Una vez creada la matriz tfidf descubrimos que tenemos 10540 palabras diferentes. sin embargo existe 5471 palabras que solo aparecen una vez lo que quiere decir que tenemos solamente y 7131 solo aparecen en dos twitts o menos.

Tras evaluar un poco mas las palabras observamos en la siguiente grafica fque las palabras mas usadas son las siguientes (eur, company, said, net, sales, million, profit, finnish). Podemos ver que estas palabras no son muy relevantes para el contexto de los sentimientos por lo que las eliminamos del corpus.

Esta grafica lo que nos indica principalment es que la mayoria de los twitts son de la zona europea enfocandose en resultados de compañias y aparentemente hay una fuerte concentracion en la economia finlandeza empresas como nokia y bancos y tecnologias moviles, en la siguiente grafica observaremos las 50 palabras mas utilizadas.

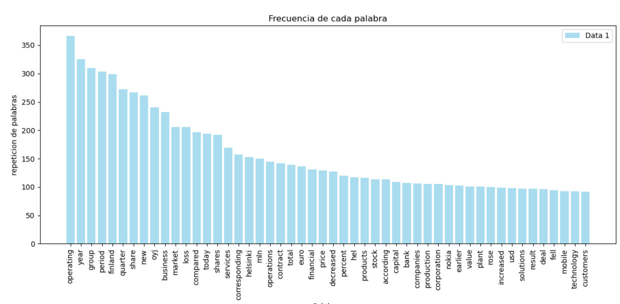

Figure 3: Distribucion de palabras Train y Test

3 Entrenamiento del modelo y selecciòn de hiperparametros

Para el entrenamiento primero generamos dos bases una utilizando gradient boosting y otra utilizando random forest, se hizo un train test split del 33%, para asegurarnos que ambos data sets sean similares adjuntamos la grafica de como estan distribuidas las frecuencias

	gradient boosting	Random forest
train	0.79	0.93
test	0.66	0.64

de las primeras 50 palabras.

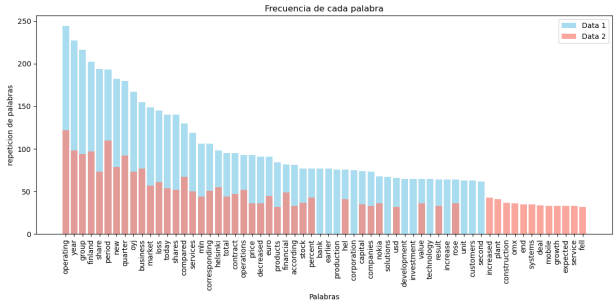

Figure 4: Enter Caption

aqui observamos que si pueden existir un poco de diferencias y por esto tendremos que tener en cuenta que hay palabras diferentes en nuestros data sets y por esto podremos perder rendimiento.

Para medir el rendimiento de nuestros modelos utilizaremos el "Accuracy" del modelo es decir que tan bien esta clasificando el modelo.

Para la linea base.

Aqui podemos observar que en los modelos de base claramente observamos un sobre ajuste de los datos al tener un valor de entrenamiento significativamente superior al de prueba, tambien podemos observar que el gradient boosting esta generalizando un poco mejor que el random forest y esta obteniendo mejores resultados en test.

Ahora bien, para obtener el mejor rendimiento utilizaremos y compararemos 2 metodos de otpimizacion de hyper parametros diferentes, el primero es un proceso de optimizacion bayesiano utilizando la liberia hyperopt, el otro metodo que utilizaremos es un metodo tradicional de random search, para estos metodos utilizamos nuevamente gradient boosting y random forest y estos fueron los resultados que obtuvimos.

Como resultado obtuvimos el resultado en la tabla anterior. podemos observar que el gradient boosting se comporto muy similar indiependiente del metodo y los hyperparametros que utilizaramos entre el hyperopt y el random search, por lo tanto el modelo elegido fue hyperopt.

Finalmente ejecutamos un modelo no supervisado para la evaluacion de topicos tipo LDA el cual nos dio como mejor resultado 3 topicos diferentes, los cuales podriamos nombrar

	Hyperopt	Random Search(RF)	Random Search (GB)
Modelo	Gradient boosting	Random Forest	Gradient Boosting
Max Depth	32	16	5
Min Sample Leaf	N/A	6	8
Min Sample Split	N/A	3	5
N_estimators	N/A	70	153
Learning rate	0.69	N/A	0.088
Final Score	0.64	0.55	0.6531

como:

1. empresas de tecnologia finlandesa
2. competencia y rendimiento de las acciones
3. rendimientos financieros generales

4 Conclusiones

Con este poryecto podemos concluir que es posible desarrollar un modelo que nos permita clasificar por sentimiento los twitts, sin embargo el rendimiento de los modelos hasta ahora no ha sido del todo satisfactorio para una puesta en produccion y por lo tanto debemos continuar con el desarrollo de este proyecto para mejorar el rendimiento, algunos pasos que se pueden tomar para mejorar el rendimiento de este ejercicio son los siguientes:

1. Eliminar las palabras que solo se aparecen una vez.
2. Explorar otros modelos.
3. eliminar letras unicas del data set.
4. balancear mejor las categorias (sentimientos).
5. aumentar la cantidad de twitts en el procesamiento del modelo.
6. resampleo de los datos.

Adicionalmente podemos concluir que el modelo mas adecuado que reduce mas el overfitting y produce un mejor resultado y mejor generalizacion es el gradient boosting, pero sin embargo existe mucho espacio para crecimiento pues una accuracy del 65% es muy bajo para llamar a este proyectyo un exito.

En el caso del Modelo LDA los topicos que identifico no fueron realmente relevantes y no aportan mayor informacion al analisis sin embargo si se obtiene informacion mas general se podria obtener informaciòn sobre aspectos macroeconomicos u otros, pero en este caso

la relevancia de los topicos no es lo suficientemente buena como para considerar que este modelo nos aporte mas informaciòn relevante.