Predicion de una Nota Puesta a un Docente Basado en una Opinión

Freddy Alejandro Cuellar Garcia
Universidad Nacional Bogotá, Colombia
Email:facuellarg@unal.edu.co

***Resumen*—La predicción de una nota basado en una opinión se puede ver como la clasificación de cada opinión en una nota, también puede verse como una regresión lineal donde el texto pasa por algún tipo de función y retorna un valor entre 0 y 5. Estos dos enfoques pueden dar resultados diferentes y dependen de la representación que le demos al texto. Dos representaciones interesantes para tratar en este problema son Word2Vect y Análisis de sentimiento (A.S.) como un arreglo de valores reales. En este documento probaremos estas dos representaciones por separado y juntas para ver cual de estas es mejor para la clasificación.**

I. Introducción.

En este documento se presentara la comparación de los resultados obtenidos usando diferentes métodos de clasificación y representación de texto. Se harán dos tipos de clasificación, una discreta y otra continua, en la discreta tendrá etiquetas que equivalen a valores $0, 0.5, 1.0, 1.5, ..., 5.0$. En la clasificación continua los posibles resultados son valores reales comprendidos entre 0 y 5. Para la representación de texto se usarán dos modelos, el modelo Word2Vect y A.S. Para los datos de entrenamiento se uso la página los estudiantes en la cual los estudiantes pueden seleccionar un profesor, darle una nota y dejar una opinión acerca de él y como es para una materia en específico. Finalmente se mostrará la comparación de los resultados usando los diferentes tipos de clasificación y representación.

II. Datos

Los datos usados para el entrenamiento de los modelos fueron extraídos de la página los estudiantes, la extracción se hizo mediante un crawler el cual recorre la página y saca la información de los profesores uno por uno, guardando la opinión y la nota dada al profesor. Las opiniones están en español, además están tal como el usuario las puso, así que pueden haber errores ortográficos los cuales no fueron tratados en es proyecto. Se recogieron un aproximado de 25000 opiniones con la respectiva nota dada. Las notas son valores entre 0 y 5 con un posible decimal.

III. Clasificador

Para la clasificación discreta se entrenó un clasificador bayesiano, con las posibles etiquetas $0, 0.5, 1.0, 1.5, ..., 5.0$ usando la herramienta *sklearn.naive_bayes* de `Python`.
Para la clasificación continua se hizo una regresión lineal usando *sklearn.linear_model.LinearRegression* de `python`.

IV. Representacion del texto

Para la representación Word2Vect se uso la herramienta *gensim.models.word2vec*, se tomó un tamaño de características igual a 100, una ventana de contexto de tamaño 5, esto para cada palabra, luego cada opinión fue representada como el promedio de la representación de cada palabra que la compone. Para la representación por A.S. se utilizó la herramienta *vaderSentiment.vaderSentiment.SentimentIntensityAnalyzer* la cual recibe la oración que se desea analizar y retorna cuatro valores, estos cuatro valores fueron los que se usaron como representación del texto.
El analizador retorna cuatro métricas:

- neg, neu, pos*: son la proporción de expresiones negativas, neutrales y positivas encontradas en el texto.
- compound*: se trata de una calificación que toma valores entre -1 y 1. Puede verse como la valencia del texto dado.

Así mismo, se puede clasificar el sentimiento del texto de acuerdo con la siguiente convención:

Sentimiento	Compound
Negativo	$compound \leq -0.5$
Neutral	$-0.5 \leq compound \leq 0.5$
Positivo	$compound \geq 0.5$

Cuadro I: salida de analizador de sentimientos

Y como última representación se uso la concatenación de las dos anteriores.

V. Metodología

En esta sección se explicara el proceso que se siguió para llevar acabo este proyecto, primero que todo se hablará del preprocesamiento que se le aplicó al texto antes de representarlo de alguna de las tres maneras ya explicadas. Se reemplazaron las tildes por la respectiva vocal sin tilde, se pasaron a minúscula todas las palabras, se quitaron caracteres especiales, menos la `ñ`, se removieron las stopwords, se separó por tokens donde cada token era una palabra y por último se eliminaron los espacios en blanco al inicio y al final de cada palabra. Para el entrenamiento los datos se separaron usando cross validation con un $fold = 4$ tomando el $75\,\%$ de datos para el entrenamiento y el $25\,\%$ para el testeo. Con la representación de análisis de sentimientos(A.S) se tuvo que hacer un poco mas de trabajo pues la herramienta *vader* solo funciona para textos en inglés, razón por la cual

se tradujeron todas las opiniones usando el API de Google Translate, debido a esto también se cambio un poco el preprocesamiento, en específico las stopwords que se usaron en este caso fueron las del idioma inglés y no el español. Luego se pasaron estos datas para el clasificador, ya fuera Bayesiano o la Regresión Lineal. Después de esto se sacaron las métricas pertinentes para cada uno.

VI. Resultados

En esta sección se mostrarán los resultados obtenidos usando estos clasificadores con las diferentes representaciones. Primero se mostrará el resultado de los r2 calculados para cada clasificación con cada representación

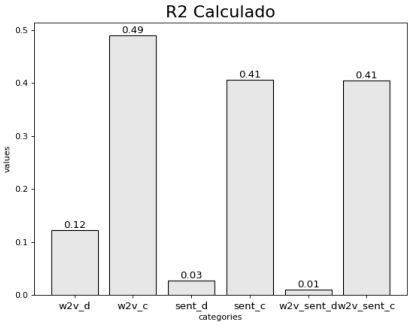

Figura 1: Calculo de r2

Como se puede ver en esta gráfica el mayor *r2* se encuentra cuando usamos wor2vect en una clasificación continua. también es notable como el juntar las dos representaciones parece no mostrar ninguna mejoría con respecto a sólo usar A.S., esto puede ser debido a que la forma correcta de juntar dichas representaciones no es simplemente concatenarlas.
Las siguientes gráficas mostrarán el conteo de las diferencias del valor predicho con respecto al real de manera proporcional.

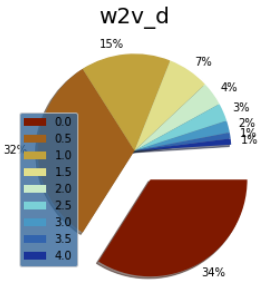

Figura 2: Diferencias usando w2v en clasificación Discreta

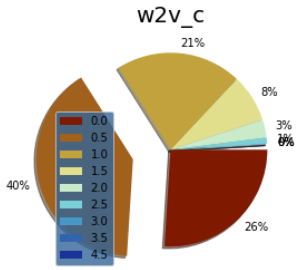

Figura 3: Diferencias usando W2V en clasificación Continua

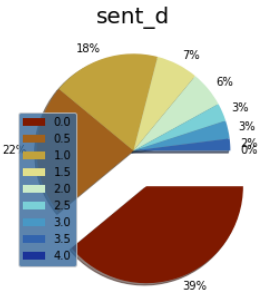

Figura 4: Diferencias usando A.S en clasificación Discreta

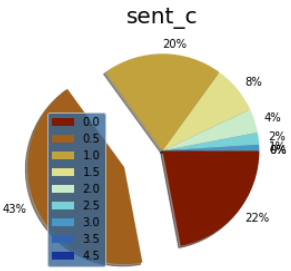

Figura 5: Diferencias usando A.S. en clasificación Continua

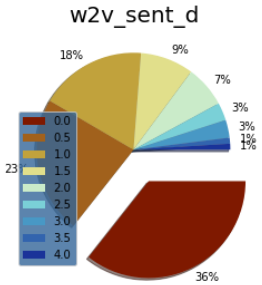

Figura 6: Diferencias usando W2V+A.S.en clasificación Discreta

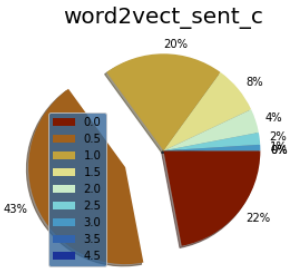

Figura 7: Diferencias usando W2V+A.S.en clasificación Continua

En todas las gráficas mostradas se ve que más del $60\,\%$ de las perdiciones hechas tienen una diferencia ≤ 0.5 con respecto a la calificación real que se le daría al profesor y el $80\,\%$ tiene una diferencia ≤ 1.

CONCLUSIONES.

Como primera conclusión se puede decir que hasta no encontrar una mejor manera de unir dos representaciones lo mejor es usar solo una. También es notorio que word2vect representa de mejor manera el texto para este problema. Y por último se puede ver que la mejor manera de abordar este problema es con una clasificación continua pues es donde se ven mejores los resultados con relación al *r2-score*

[1]

[1] https://github.com/facuellarg/reviews_nlp