Comparación entre Árbol de Decisión y Clasificación por Léxico aplicado a PNL

Carlos Velazquez

Facultad Politécnica – U.N.A., Campus, San Lorenzo, Paraguay
secretaria@pol.una.py
https://www.pol.una.py/institucional/

Resumen El crecimiento constante de publicaciones realizadas en las redes sociales ha dejado abierta la posibilidad de obtener una gran cantidad de información acerca de las opiniones realizadas sobre los más diversos temas. Actualmente se cuenta con varios algoritmos capaces de aprovechar este nuevo flujo de información, sin embargo, convertir esos datos en información precisa todavía es cuestión de estudio.
En base a publicaciones realizadas en Twitter, este reporte ha tomado los algoritmos de árbol de decisión y clasificación por léxico y los ha aplicado sobre el mismo conjunto de datos para determinar el sentimiento que genera cada publicación. Una vez obtenida la predicción se ha buscado comparar los resultados con diversidad métricas para determinar como el conjunto de inicial de datos puede afectar el resultado y como poder mejorarlos con el fin de obtener mejores predicciones.

Keywords: Procesamiento Lenguaje Natural · Árbol de decisión · Clasificación por Léxico .

1. Introducción

El uso de un conjunto de datos provenientes de una red social puede ser difícil de manipular y procesar. El objetivo de este estudio es determinar la manera más efectiva de tratar con estos datos y cuáles son los puntos que pueden mejorarse para obtener predicciones precisas.

2. Metodología

2.1. Comparaı¿ción de Resultados entre el Árbol de deCisión y Clasificación por Léxico

Al aplicar diferentes algoritmos sobre un mismo conjunto de datos, se busca determinar como la entrada inicial puede influir en la salida de datos. Este resultado será medido finalmente utilizando como métrica la presición alcanzada comparando los valores predichos con el grupo de control. Este grupo de control es definido usando el sentimiento que ya fue asociado a cada tweet por una persona.

Para la Clasificación por Léxico se ha utilizado un listado de palabras positivas y otra lista para las palabras negativas.

2.2. El Conjunto Inicial de Datos

Este caso de estudio se centra en publicaciones realizadas en Twitter sobre el COVID-19 para determinar el nivel de precisión a la hora de realizar predicciones con varios algoritmos. Sobre estos datos se aplicarán varios niveles de pre-procesamiento como eliminar palabras no relevantes para el análisis, agrupamiento por raíces de palabras y remover signos de puntuación.

Pasos del pre-procesamiento:

- Case Converter: Como las palabras mayúsculas y minúsculas se consideran diferentes palabras, se pasaron todas a minúsculas. De esta manera se facilita su agrupación y obtención de frecuencias y terms.
- Cell Splitter: Como cada texto de la publicación contiene su id, que no es relevante, se separó este dato en una nueva columna y el resto del texto en otra columna.
- String to document: paso necesario en Knime para procesar los datos. Para la obtención final del documento se utilizó la columna título generada en el paso anterior y para el texto se utilizó la columna texto que también fue generado en el paso anterior. Además se utilizó la columna sentimiento como categoría del documento.
- Column Filter: Como de aquí en adelante solo se utilizó la columna document generada en el paso anterior, todas las demás columnas fueron removidas.
- Stop word filter: Removidas las palabras consideradas como poco relevantes para la clasificación del documento. Para este paso se instaló una extensión del Knime para obtener un listado de palabras consideras como poco relevantes.
- Snowball Stemmer: Utilizado para obtener los términos de las palabras. También se utilizó una lista para el idioma español en este paso.

3. Resultados

De los 4800 tweets, 1648 fueron clasificados mediante Léxicos. El porcentaje final de precisión fue de 41 %. De los mismos 4800 tweets, se predijeron resultados para 960, quedando el resto como datos de entrenamiento. Estas predicciones arrojaron un porcentaje de precisión del 51 %.

Cuadro 1. Métricas obtenidas para el Árbol de decisión.

Sentimiento	True Pos	False Pos.	True Neg.	False Neg	Recall	Precisión	Accuracy
NEUTRO	211	203	386	157	0.573	0.506	-
NEGATIVO	204	172	4214	163	0.556	0.543	-
POSITIVO	75	92	643	150	0.333	0.449	-
OVERALL	-	-	-	-	-	-	0.51

Cuadro 2. Métricas obtenidas para Clasificación por Léxico.

Sentimiento	True Pos	False Pos.	True Neg.	False Neg	Recall	Precisión	Accuracy
NEUTRO	364	362	643	279	0.566	0.501	-
NEGATIVO	261	128	665	594	0.0305	0.671	-
POSITIVO	51	482	1016	99	0.34	0.096	-
OVERALL	-	-	-	-	-	-	0.41

4. Conclusiones

La precisión alcanzada por ambos algoritmos tiene una diferencia notable del 10 %, sin embargo, ambas son bajas ya que están por debajo del 80 % aceptado normalmente [3] para análisis de sentimiento. Esta baja precisión con ambos métodos demuestra que es la entrada de datos la que puede ser mejorada para obtener mejores resultados.

Algunos puntos a mejorar son:

- Limpieza de palabras no relevantes: Los enlaces y los usuarios tageados son muy poco relevantes a la hora de predicir el sentimiento [1].
- Limpieza de signos de exclamación: En algunos tweets se abusan de los signos de exclamación por lo que eliminar estos datos podría mejorar la precisión.
- Diccionario de palabras: Esto solo se aplica a la clasificación por léxico, pero un listado más completo definitivamente ayudará a obtener mejores resultados.

Punto fuerte:

- Los hashtags: Estos elementos demostraron ser bastante relevantes a la hora de hacer predicciones. Algunos como #quedateencasa tuvieron un gran impacto en las clases.

Referencias

1. Chiorrini, Andrea; Diamantini, Claudia; Mircoli, Alex; Potena, Domenico. .Emotion and sentiment analysis of tweets using BERT"(PDF). Proceedings of Data Analytics solutions for Real-LIfe APplications (DARLI-AP) (2021)
2. Hartmann, Jochen ; Heitmann, Mark; Siebert, Christian; Schamp, Christina. "More than a Feeling: Accuracy and Application of Sentiment Analysis".International Journal of Research in Marketing (2022)
3. Lexalytics, Sentiment Accuracy: Explaining the Baseline and How to Test It . Último acceso 20 dec 2022