Universidad de los Andes

Departamento de Ingeniera de Sistemas Y Computación

Creación de modelo predictivo para clasificación de Tweets con connotación depresiva

Proyecto Inteligencia de Negocios
ISIS3301

Autores:

Martin David Galvan Castro -201614423
Juan Camilo Sánchez -201519337

Abril 2020

Contenido

Tabla de Contenidos

1. Comprensión del Negocio y los Requerimientos

Tabla 1. Comprensión del negocio

Oportunidad/Problema del Negocio	Se creo una red social de apoyo psicológico que se basa en usuarios médicos y usuarios pacientes que ofrecen una ayuda inicial a los pacientes. Esto para conectar médicos y pacientes para empezar una relación beneficiosa para ambos. Este proyecto está buscando obtener publicidad orientada a los clientes para que se subscriban a la red social	
Descripción del Requerimiento desde el punto de vista de minería de datos	Para obtener los usuarios a los cuales enviarles publicidad se decidió buscar en twitter por lo cual se necesita identificar aquellos perfiles que hayan realizado tweets depresivos que probablemente apreciarían hablar sobre sus problemas. Así se deben descartar los que no tengan nada que ver y obtener los relacionados.	
Detalles de la actividad de minería de datos		
Tarea	Técnica	Algoritmo y parámetros utilizados
Supervisada	De Clasificación	Basado en arboles de decisión

2. Comprensión de los datos y preparación

Para entender los datos, la primera aproximación que se hizo fue una inspección de los datos. Al finalizar la inspección de los datos se pudo encontrar que los datos contenían tres variables: el ID, el tweet y la asignación de la clase. A partir de esta inspección se planeó un flujo de trabajo para filtrar el contenido dentro de los tweets. Se observo que algunos tweets contenían emojis, menciones a otros usuarios, eran re-tweets de otro usuario o uso de caracteres especiales.

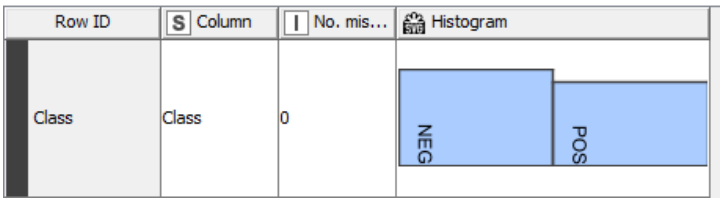

Ilustración 1.Perfilamiento datos de clase

Además de esto, no se encontró ningún dato faltante en la variable correspondiente a la asignación de la clase. A continuación, se puede ver un ejemplo de cómo son los datos sin hacer el correcto preprocesamiento y un perfilamiento de la variable *target:*

Id	Text	Target
289612631038361...	Argh.. I hate my life	0
289612727654170...	I'm good	1
289612736063758...	Enjoy my life	1
289612773716008...	RT : I do what I want	1
289612819807211...	My life is just a series of unfortunate fucked up events @margaretkuta @mwrighht	0
289612853529432...	Im tryna become better everyday I open my eyes.... Starting to really understand LIFE IS AMAZING	1
289613134543613...	RT : I hate how much I over-think at night	0
289613188968894...	RT @sydthekid_15: I just want to be genuinely happy	0
289613243461296...	My life is one big mess	0
289613285307863...	oh I have lol I just eventually get what I want though	1
289613381936242...	RT : I hate how much I over-think at night	0
289613490908434...	I swear this fucking shit for me	0
289613490937798...	yup i hate my life	0
289613587440340...	RT @_misterCT Being single Sucks... I ain't bout this life no more _	0
289613608353140...	I have the worst headache ever imaginable	0
289613621082865...	Why don't I have a life	0
289613680049115...	WHY DO I SUCK AT LIFE	0
289613713433042...	Having a broken finger for the rest of my life sucks	0
289613717405065...	I wish I can say I am happy	0
289613801458921...	I do what i want to	1
289613813836283...	I wanted a change, lil nigga	0
289613826498887...	I wish.... I was different	0
289613847483015...	RT : My life is perfectly terrible	0
289613939736707...	Soo happy! With my life with you with everything!_ #couldntaskformore	1
289613951212322...	RT : i do what i want	1
289613956618780...	I love how I can only trust 2 people in my life❤ that's all I need	1
289614019604660...	RT @Libras_R_Us: As a #Libra, I get revenge by living a great life	1
289614095198593...	RT : To stressed to see that I'm blessed	1
289614103373295...	I've never been so disappointed in my life	0
289614111791263...	I'm such a GOOD PERSON ❤❤ ! #SeriousTweet	1

Ilustración 2. Ejemplo de datos sin filtrar

Dentro de este flujo primero se contempla hacer un reemplazo en la variable de asignación de la clase, ya que estos están representados con 1 para un tweet sin connotación depresiva y 0 para un tweet con connotación depresiva. Esto se reemplazó por un POS si es 0 y NEG si es 1. Seguido a esto se hicieron las siguientes operaciones, esto con el fin de que al tener el formato final de las bolsas de palabras, solo se tengan en cuenta palabras claves que se repitan con mucha frecuencia en los documentos:

- Filtrar columnas innecesarias, ya que ID no se usa
- Convertir el texto en un objeto documento, para ser proceso
- Eliminar menciones, se hizo con una expresión regular de la forma \@.*
- Convertir caracteres especiales, puntuación y palabras con menos de 3 caracteres, para obtener las palabras claves por documento
- Eliminar Stopwords, para eliminar palabras innecesarias
- Hacer lowercase a los textos, para tener todas las palabras en el mismo formato
- Lematizar el texto, para evitar plurales y conjugaciones de palabras
- Crear bolsa de palabras
- Filtrar los términos obtenidos de acuerdo con la frecuencia en los documentos, para evitar que palabras que no son frecuentadas, como nombres, links, hashtags no sean tenidos en cuenta
- Crear un bit array para cada documento, para poderlo usar dentro del clasificador

Al aplicar estas transformaciones, resultamos con una matriz de la siguiente forma:

Document	feel	love	life	live	happi	bless	chang	shit	suck	fuck	D
"HAHAHAH my life sucks"	0	0	1	0	0	0	0	0	1	0	0
"wouldn't change a thing in my life❤"	0	0	0	0	0	0	1	0	0	0	0
"words cannot describe how happy i am"	0	0	0	0	1	0	0	0	0	0	0
"woehoe 25252 love my life"	0	1	1	0	0	0	0	0	0	0	0
"woah ~ I'm in love"	0	1	0	0	0	0	0	0	0	0	0
"wo! I have the worst headache ever"	0	0	0	0	0	0	0	0	0	0	0
"wish so many things in my life were different #wishfulthinking"	0	0	1	0	0	0	0	0	0	0	0
"wish my life was different"	0	0	1	0	0	0	0	0	0	0	0
"wish i died..or,like..i move to a different school.naa..wish i was dead"	0	0	0	0	0	0	0	0	0	0	0
"wise words, I'm proud"	0	0	0	0	0	0	0	0	0	0	0
"winter breaks....i have no life"	0	0	1	0	0	0	0	0	0	0	0
"HAHAHA I hate my life"	0	0	1	0	0	0	0	0	0	0	0
"why you're sad? — I'm not sad, i'm always happy(: http://t.co/W6lnjOFt"	0	0	0	0	1	0	0	0	0	0	0
"why is my life so stupid"	0	0	1	0	0	0	0	0	0	0	0
"why is my life so difficult"	0	0	1	0	0	0	0	0	0	0	0
"why is my life so boring fml"	0	0	1	0	0	0	0	0	0	0	0
"why is my life so boring"	0	0	1	0	0	0	0	0	0	0	0
"why i hate my life so much"	0	0	1	0	0	0	0	0	0	0	0
"why does my life have to suck so bad"	0	0	1	0	0	0	0	0	1	0	0

Ilustración 3. Resultado final de los datos

3. Modelado y Evaluación

Para el modelado se empleó un árbol de decisión. Para esto se implementaron tres versiones:

- Árbol de decisión con partición de datos: Se hace una partición de los datos de manera que se define el 70% de los datos para entrenamiento y el 30 restante para validación.
- Árbol de decisión con partición de datos y SMOTE: Se hace la misma aproximación que en el numeral anterior, pero se aplica el algoritmo SMOTE para generar datos en la clase minoritaria
- Árbol de decisión con validación cruzada: Se aplica validación cruzada al árbol de decisión, se hacen 20 iteraciones. Este número de iteraciones se eligió ya que era el que mejor resultados daba.

Se decidió hacer esto a manera de verificar que los resultados obtenidos por el modelo de clasificación dieran un resultado consistente entre diferentes implementaciones, y determinar cuál de estas implementaciones era el que mayor precisión. A continuación, se mostrará el árbol obtenido, la matriz de confusión y los evaluadores correspondientes.

Para la correcta interpretación de estos árboles, se debe tener en cuenta que las variables están categorizadas como 1 o 0, por lo que cuando se hace referencia en el árbol que un término es mayor a 0.5 es porque tiene un valor de 1, que es presente en documento y si es menor a 0.5 es por que tiene un valor de 0, que es no presente en el documento.

3.1. Árbol de decisión con partición de datos

A continuación, se muestran los 3 primeros niveles de este árbol de decisión:

Ilustración 4. árbol de decisión #1

La matriz de decisión de este árbol de decisión nos muestra que tiene una precisión del 85%, que para la gran variedad de cómo están estructurados los documentos se encuentra alta. Con respecto a los evaluadores, se tiene que se obtuvo un F-Score aproximado de 0.858, por lo que se puede determinar que esta fue una buena primera aproximación.

Document ...	NEG	POS
NEG	1227	220
POS	187	1228

Correct classified: Wrong classified: 407

Accuracy: 85,779 % Error: 14,221 %

Cohen's kappa (κ)

Ilustración 5. Matriz de confusión árbol #1

3.2. Árbol de decisión con partición de datos y SMOTE

A continuación, se muestran los 3 primeros niveles de este árbol de decisión:

Ilustración 6. árbol de decisión #2

La matriz de decisión de este árbol de decisión nos muestra que tiene una precisión del 85%, se encuentra que se obtuvo aproximadamente la misma presión que el método anterior. Con respecto a los evaluadores, se tiene que se obtuvo un F-Score aproximado de 0.851, que sigue siendo muy cercano a la implementación anterior, y es indicativo de que estos modelos aproximados pueden ser uno de los mejores modelos que se pueden obtener en primera instancia.

Document ...	NEG	POS
NEG	1224	222
POS	205	1242

Correct classified: Wrong classified: 427

Accuracy: 85,24 % Error: 14,76 %

Cohen's kappa (κ)

Ilustración 7. Matriz de confusión árbol #2

3.3. Árbol de decisión con validación cruzada

A continuación, se muestran los 3 primeros niveles de este árbol de decisión:

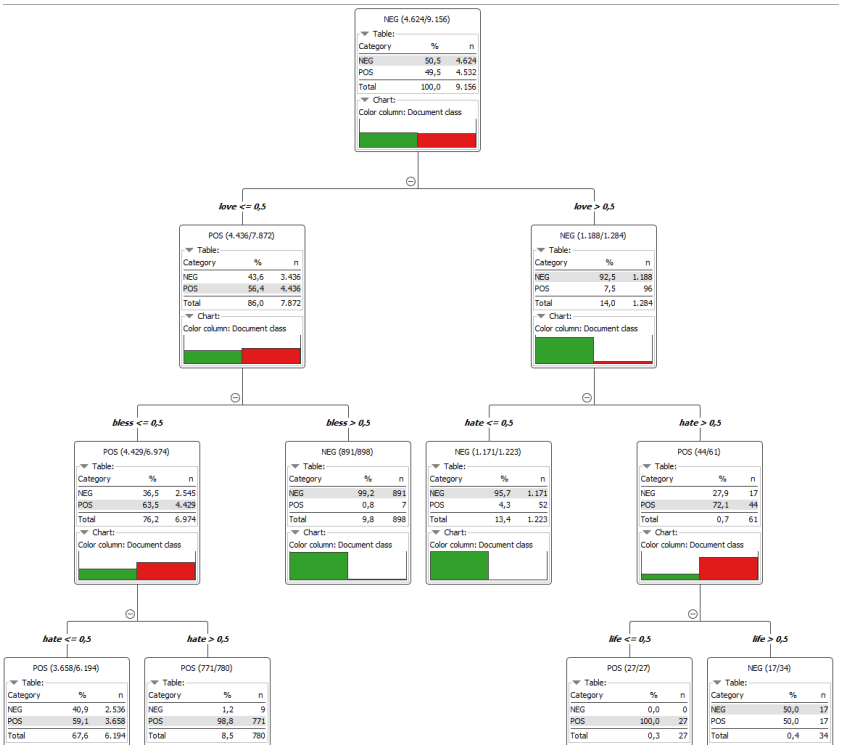

Ilustración 7. árbol de decisión #3

La matriz de decisión de este árbol de decisión nos muestra que tiene una precisión del 86%, esta fue la mayor precisión obtenida pero la diferencia no es significativa con respecto a las anteriores implementaciones. Con respecto a los evaluadores, se tiene que se obtuvo un F-Score aproximado de 0.861, que sigue siendo muy cercano a las implementaciones anteriores. Con el resultado de este modelo, se determinó que el modelo actual es el mejor posible. Por lo que para la implementación en el negocio se recomendaría usar este modelo.

Document ...	NEG	POS
NEG	4133	688
POS	650	4066

Correct classified: Wrong classified:

Accuracy: 85,97 % Error: 14,03 %

Cohen's kappa (κ)

Ilustración 8. Matriz de confusión árbol #3

4. Análisis de Resultados y Conclusiones

Según los resultados obtenidos, se determino que el mejor modelo es el de validación cruzada, aunque no se tiene una diferencia significativa con los anteriores modelos, este fue el que mayor precisión obtuvo. Adicionalmente, al ver los árboles de decisión obtenidos, se puede ver que hay palabras que son claves en la decisión en los nodos, por lo que se puede afirmar que estas palabras hacen parte de un conjunto de *Keywords* que deben ser tenidas en cuenta por el negocio en la implementación del modelo. Finalmente, se puede determinar que se hace con éxito una clasificación de tweets con connotación depresiva, ya que se obtuvo un modelo con una precisión alta, que es del 85%

5. Descripción del Trabajo en Equipo

El trabajo en equipo es una cualidad importante al momento de trabajar en proyectos de tan alta complejidad. Para la realización de este proyecto se planteó el siguiente diagrama de Gantt para hacer la repartición de tareas. Adicionalmente, en la repartición de puntos, se decidió repartirlos de manera equitativa entre los integrantes, ya que, aunque la implementación del modelo en Knime represento ser uno de los mayores retos, esta solo fue posible gracias a la extensa investigación de los miembros del equipo para determinar como seria la mejor manera de procesar los textos para obtener un modelo lo más preciso posible. Adicionalmente, en la parte de presentación de resultados se definió que sería más prudente que una persona se encargara de esto, ya que ayudaría a ahorrar tiempo en la elaboración del documento y de la presentación.

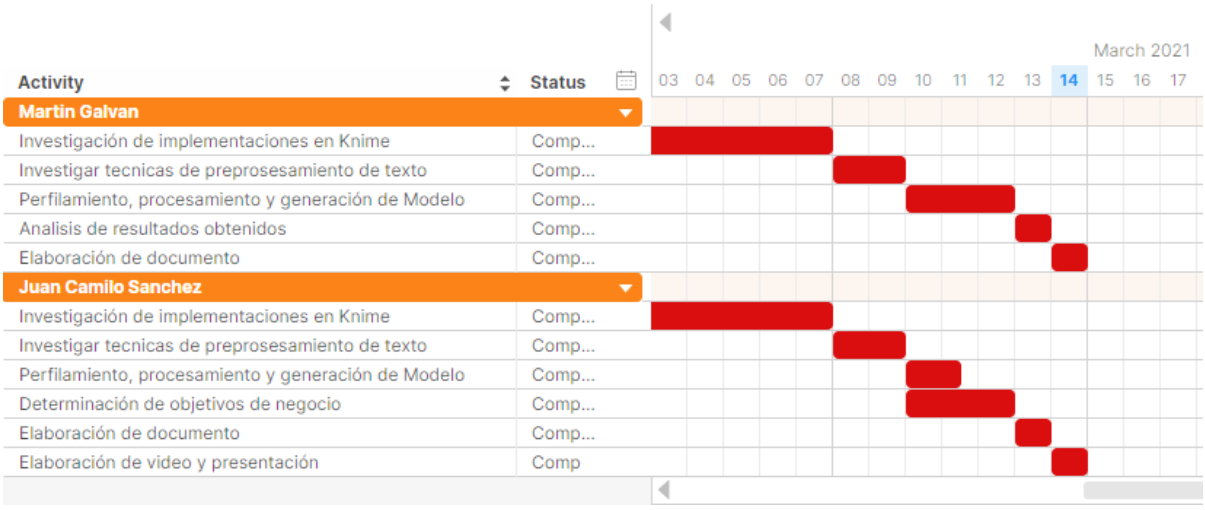

Ilustración 7. Diagrama de Gantt